GUIDE TO THE ANALYSIS OF

THE WORKPLACE EMPLOYEE RELATIONS SURVEY 1998

[Version 1.1: April 2000]

John Forth & Simon Kirby

WERS98 Data Dissemination Service

National Institute
of Economic and
Social Research

2 Dean Trench Street, Smith Square, London SW1P 3HE

Tel: +44(0) 20 7654 1902 E-mail: wers98@niesr.ac.uk
URL: http://www.niesr.ac.uk/niesr/wers98

Contents

1. INTRODUCTION 5

1.1 THE 1998 WORKPLACE EMPLOYEE RELATIONS SURVEY 5
1.2 THE CONTENT OF THE GUIDE 5
1.3 NOTATION USED IN THIS GUIDE 6
1.4 FURTHER INFORMATION 6

2. NECESSARY PREPARATION BEFORE BEGINNING YOUR ANALYSIS 9

2.1 WERS98 USER GUIDE AND VARIABLE NOTES 9
2.2 STATA MEMORY ALLOCATION 9

3. FINDING YOUR WAY AROUND THE WERS98 DATA FILES 11

3.1 WEIGHTED AND UNWEIGHTED DATA FILES 11
3.2 VARIABLE NAMING CONVENTIONS 11
*3.2.1 Variables in **Mq98fin.***: 11*
*3.2.2 Variables in **Wrq98.***: 12*
*3.2.3 Variables in **Seq98.***: 12*
*3.2.4 Variables in **Pq_9098.*** & **Pq_98out.***: 12*
3.3 THE LAYOUT OF THE DATA FILES 13
*3.3.1 **Mq98fin.*** and **Wrq98.***: 13*
*3.3.2 **Seq98.***: 14*
*3.3.3 **Pq_9098.*** & **Pq_98out.***: 14*
3.3.4 Restricted data files: 15
3.3.5 Final note: 16

4. WEIGHTING 17

4.1 THE 1998 CROSS-SECTION DATA (MANAGERS AND WORKER REPS) 17
4.1.1 Principles of weighting the 1998 data from managers and worker reps 17
4.1.2 Weight variables to be used in analysis of 1998 data from Managers and Worker Reps 19
4.1.3 A practical example of the difference between weighting schemes 19
4.2 THE 1998 CROSS-SECTION DATA (EMPLOYEES) 20
4.2.1 Principles of weighting the 1998 data from employees 20
4.2.2 Weight variable to be used in analysis of 1998 data from employees 21
4.3 THE 1990-98 PANEL DATA (PQ_9098.*) 21
4.3.1 Principles of weighting the 1990-98 panel data 22
4.3.2 Weight variable to be used in analysis of the 1990-98 panel data 22
4.4 THE 1998 OUTCOMES DATA (PQ_98OUT.*) 22
4.4.1 Principles of weighting the 1998 outcomes data 23
4.4.2 Weight variable to be used in analysis of the 1998 outcomes data 23
4.5 APPLYING AND REMOVING WEIGHTS 23
4.5.1 Applying and removing weights within SPSS 23
4.5.2 Applying and removing weights within STATA 24
4.6 THE IMPLICATIONS OF SAMPLE DESIGN FOR STATISTICAL INFERENCE 26
4.6.1 Frequency analysis 29
4.6.2 Tabular analysis 30
4.6.3 Regression analysis 32

5. THE PRODUCTION OF HIGH QUALITY TABLES IN SPSS 37

5.1 INTRODUCTION 37
5.2 PREPARATION 37
5.3 BASIC TABLE SPECIFICATION 38
5.4 MORE COMPLEX SPECIFICATIONS 41
5.4.1 Summarising continuous variables 41
5.4.2 Aggregating continuous variables 42
5.4.3 Multiple-response items 43
5.5 FINAL NOTES 45

6. COMBINING DATA FROM SEPARATE FILES FOR LINKED ANALYSIS 47

6.1 COMBINING DATA FROM THE MANAGEMENT AND WORKER REPRESENTATIVE DATA FILES 47
6.1.1 Combining the data in SPSS 49
6.1.2 Combining the data in STATA 50
6.2 ADDING WORKPLACE DATA TO THE SURVEY OF EMPLOYEES DATA FILE 50
6.2.1 Adding the workplace data in SPSS 51
6.2.2 Adding the workplace data in STATA 52
6.3 AGGREGATING DATA FROM THE SURVEY OF EMPLOYEES 53
6.3.1 Aggregating data from the Survey of Employees in SPSS 53
6.3.2 Aggregating data from the Survey of Employees in STATA 56
6.3.3 A note about the generalizability of aggregated data from the Survey of Employees 57
6.4 COMBINING INTERVIEW DATA WITH VERBATIM TEXT 61
6.4.1 The spreadsheets of verbatim text 61
6.4.2 How to export data from a spreadsheet for use in SPSS or STATA 62
6.4.3 How to export data from SPSS or STATA and add it to a spreadsheet 64

7. ACKNOWLEDGING THE USE OF THE WERS98 DATA IN PUBLICATIONS 67

7.1 ACKNOWLEDGEMENT AND DISCLAIMER 67
7.2 BIBLIOGRAPHIC CITATION 67
7.3 DEPOSITING COPIES OF PUBLICATIONS AND DERIVED DATA SETS 67

8. THE WIRS BIBLIOGRAPHY 69

APPENDIX A: LIST OF WERS98 DATA FILES AND DOCUMENTATION 71

APPENDIX B: CONTACTING THE DATA ARCHIVE: 77

APPENDIX C: INSTITUTIONS PROVIDING SHORT COURSES ON THE ANALYSIS OF SURVEY DATA USING SPSS OR STATA 79

APPENDIX D: CONTACT DETAILS FOR THE WERS98 DATA DISSEMINATION SERVICE 83

APPENDIX E: OUTPUT FROM THE SPSS TABLES MODULE 85

REFERENCES 89

1. Introduction

1.1 The 1998 Workplace Employee Relations Survey

The 1998 Workplace Employee Relations Survey (WERS98) is the fourth in an internationally regarded series in which key role-holders provide extensive information on the nature of employment relations at their place of work. The first survey in the series was conducted in 1980; subsequent surveys also took place in 1984 and 1990.

The principal component of each survey in the series is a face-to-face interview at the establishment with the senior person dealing with industrial relations, employee relations or personnel matters. Interviews are also sought with worker representatives, where present. These two elements form the core of the four cross-section surveys in the series. The 1998 cross-section survey was, however, the first in the series to include a survey of employees. WERS98 also included a more extensive panel survey than had been attempted in previous years.

Developments in the methodology of the survey were accompanied by changes in the content of the interview schedules used in the cross-section and panel surveys. New topics in the cross-section management interview included equal opportunities, flexible working practices and management attitudes. The panel survey, for its part, was the first to use an interview schedule specifically designed to investigate change.

These various innovations will have attracted many analysts with no previous experience of using data from the series. However, innovations in the design of the 1998 survey will also mean that analysts with much experience of using data from previous surveys in the series will also inevitably be faced with new challenges. The aim of this Guide is to provide both the new and the experienced user with some assistance as they begin to analyse the wealth of data available from WERS98.

1.2 The content of the Guide

The Guide aims to cover the most common issues that will face the user in their analysis of WERS98. Its content ranges from the production of simple tables to the use of weighting in multivariate analysis, and it is designed to be of use to both experienced and inexperienced analysts.

The Guide focuses primarily on analysis of the WERS98 data using SPSS 9.0 for Windows and Intercooled STATA 6.0 for Windows. We have chosen to concentrate on SPSS and STATA since these are the formats in which most users will access the data. However, the WERS98 data is also available in SAS and ASCII formats.

The Guide contains many practical examples and assumes that users have access to the SPSS/STATA data files and all of the associated documentation. A full list of the available data files is given in Tables 1 and 2 of Appendix A; the full range of documentation that accompanies the survey data is listed in Tables 4 and 5. Each of the data files may be obtained from the Data Archive at the University of Essex (see Appendix B). The documentation is available in electronic form on the web-sites of

both the Data Archive (Appendix B) and the WERS98 Data Dissemination Service (see Appendix D), whilst the Data Archive can supply hard copies for a small charge.

The practical guidance given assumes that each of the relevant data files is stored on the users' hard disk in a directory named D:\WERS98\. Those using a different storage mechanism or directory path will need to amend the syntax or menu instructions accordingly.

Readers using SPSS 9.0 for Windows should note that the procedures required to complete each of the practical examples outlined in the Guide are given in both syntax and menu-based format. Menu options, in particular, may differ in earlier versions of SPSS.

Finally, the reader should please note that this Guide is intended to cover analytical issues that are particular to the analysis of WERS98. It is not intended as a general guide to the operation of SPSS or STATA, nor to the general principles of survey analysis. Short courses covering these general topics are regularly available from the institutions listed in Appendix C. In addition, both SPSS and STATA come with on-line help systems and on-line tutorials.

1.3 Notation used in this Guide

There are a small number of conventions that have been adopted throughout this Guide. These are as follows:

- Variable names appear in bold, capitalized font (e.g. **ASTATUS**)
- Names of data files appear in bold, lower case font (e.g. **Mq98fin.***). Often an asterisk (*) is used in place of a particular suffix when the point being made in the text applies to files irrespective of their format.
- References to specific SPSS or STATA commands appear in courier font (e.g. `weight by est_wt`).

In addition, since the first three surveys in the series were named the Workplace Industrial Relations Surveys, for ease, we retain the former acronym in this Guide when referring to the series as a whole (the WIRS series). We use the new acronym (WERS98) when referring specifically to the most recent survey.

1.4 Further information

Users wishing to consult the primary analyses of the WERS98 data are referred to three volumes:

Cully M, Woodland S, O'Reilly A, Dix G, Millward N, Bryson A and Forth J (1998) *The 1998 Workplace Employee Relations Survey: First Findings*, London: Department of Trade & Industry. (ISBN: 0-856-05382-1)

A 30-page booklet of initial findings from the survey, published in October 1998. Available free of charge from the Department of Trade & Industry. Telephone the DTI Publications Order Line on

+44 (0)870 1502 500, quoting the title and reference number (URN 98/934), or download the document from the DTI web site.[1]

Cully, M., Woodland, S., O'Reilly, A. and Dix, G. (1999) *Britain at Work: As Depicted by the 1998 Workplace Employee Relations Survey*, London: Routledge.

Contains a full and detailed primary analysis of WERS98. Published in September 1999, this 341-page volume constitutes the principle volume of findings from the 1998 Survey. Priced £20 paperback (ISBN: 0-415-20637-5); £60 hardback (ISBN: 0-415-20636-7). Copies may be ordered direct from Routledge, by telephoning +44 (0)1264 342939.

Millward, N., Bryson, A. and Forth, J. (2000, forthcoming) *All Change at Work? British Employment Relations 1980-98 as portrayed by the Workplace Industrial Relations Survey Series*, London: Routledge.

Companion volume to *Britain at Work*, focusing on change over the course of the Survey series. Makes extensive use of each of the four cross-section surveys of 1980, 1984, 1990 and 1998, together with the 1990-98 panel survey. Also priced £20 paperback (ISBN: 0-415-20635-9); £60 hardback (ISBN: 0-415-20634-0) and available from Routledge. Scheduled publication date: 12 May 2000.

Further information about the 1998 Workplace Employee Relations Survey is available on the web-site of the WERS98 Data Dissemination Service (see Appendix D), from where users may also view or download an electronic version of this *Guide to Analysis*.

[1] URL: http://www.dti.gov.uk/IR/emar/ffind.pdf [verified: 10/4/0].

2. Necessary preparation before beginning your analysis

2.1 WERS98 User Guide and Variable Notes

Before beginning to analyse the WERS98 data, users should ensure that they are familiar with those elements of the User Guide that are relevant to the particular data set they intend to work with.

Users should also ensure that they have consulted the set of *Variable Notes* that has been produced to accompany each of the WERS98 datasets (see Table 5 in Appendix A). These *Variable Notes* list all known variable-specific issues that may be of interest to the analyst when using the data from WERS98. Such problems might range from small errors in the description of a filter in the questionnaire to more fundamental problems in the operation of a particular question within the interview. Consulting these *Variable Notes* before starting work could save considerable amounts of time and effort spent investigating issues already resolved by other users.

We rely upon users to assist us in keeping these *Variable Notes* to date. We therefore request that all users notify the Data Dissemination Service of any new problems that they discover in either the data files or documentation during the course of their work.

Information will be posted on the Data Dissemination Service web site at regular intervals to notify users of new data and documentation as they become available. Users that have registered with the Data Dissemination Service will automatically receive notification of updates to the web site by e-mail. The WERS98 Data Dissemination Service web-site contains details of how to register (follow the link to 'Contacting the WERS98 Data Dissemination Service').

2.2 STATA memory allocation

By default, STATA allocates 1,000 kilobytes (1 Mb) of memory space for you to work with. This memory space is used to store data and run procedures. Hence, you must ensure that the memory space is large enough to both store your data file and run the analyses that you want to conduct on it.

The STATA versions of the WERS98 Cross-section data files on general release have the following sizes:

Main Management data file (**Mq98fin.dta**)	2,568 Kilobytes (2.57 Mb)
Worker Rep data file (**Wrq98.dta**)	345 Kilobytes
Survey of Employees data file (**Seq98.dta**)	2,882 Kilobytes

The STATA versions of the WERS98 Panel Survey data files on general release have been divided up so as to comply with STATA's limitations on the maximum number of variables permitted within a single file. The separated files have the following sizes:

Panel Interview data, comprising:

1990 management data (**Pq_9098a.dta**)	1,128 Kb
1990 worker rep and financial manager data (**Pq_9098b.dta**)	1,146 Kb
1998 management data (**Pq_9098c.dta**)	1,482 Kb

Panel outcomes data, comprising:

1990 management data and 1998 outcome code (**Pq_98outa.dta**)	2,440 Kb
1990 worker rep and financial manager data (**Pq_98outb.dta**)	3,248 Kb

Studying this information, one can see that only the Worker Rep data file is small enough to be opened under the default memory setting of 1,000 kilobytes. Around 640 Kb are left for STATA to work with after opening this file.

STATA's memory allocation can be increased, either for the purposes of opening the larger files or for running complex procedures, by using the `set memory` command. This command works in Kilobytes, so to increase the memory allocation to 5,000 kilobytes (5 Mb), for example, one would first clear the memory of all data (using the `clear` command) and then type:

```
set memory 5000
```

For further information, see Chapter 7 of the STATA User Guide.

3. Finding your way around the WERS98 data files

The WERS98 data files have some particular features that it is useful to be aware of at the beginning of your analysis.

3.1 Weighted and unweighted data files

Users should note that some (but not all) of the SPSS versions of the WERS98 data files on general release have been saved with the weight already applied to the data. This means that they are ready to produce weighted analyses as soon as they are opened in SPSS. These files are:

- **Wrq98.por**
- **Seq98.por**

To produce unweighted analyses of the data contained in these files, the user must first remove the weighting from the data. See Section 4: Weighting.

All other SPSS data files, and all files in other formats such as STATA, are supplied unweighted. In order to produce weighted analyses from these files, the user must apply the weight to the data. Again, see Section 4: Weighting.

Users can independently establish whether a particular data file has been saved in weighted form by examining the SPSS Data Editor (similar in appearance to one page of a spreadsheet). With a data file open in the Data Editor, the user should look to the bottom right-hand corner of the screen. If the data is weighted, the phrase 'Weight on' will appear in one of the boxes adjacent to that containing the phrase 'SPSS for Windows Processor is ready'. If 'Weight on' is not present, the data is currently unweighted.

3.2 Variable naming conventions

All variable names used in the WERS98 data files are no more than 8 characters in length. In general, each variable name has two parts: a one or two-character prefix that signifies which section of the relevant questionnaire the variable arises from; and a remainder of up to seven characters that is intended to give some sense of the topic covered by the question. Variables arising from questions that permitted multiple responses have a number at the end to signify the order of response.

*3.2.1 Variables in **Mq98fin.**:*

A one-character prefix signifies the section of the Main Management questionnaire from which the variable arises. So **ASTATUS** arises from Section A of the questionnaire. Variables arising from multiple response questions are numbered from 1 upwards (or, from 01 if 10 or more responses were permitted), so that **AHOWCHA1** contains the first numeric response given by a particular manager to the question about changes of ownership, and **AHOWCHA7** the seventh response. Note, however, that few respondents gave the maximum number of responses to any

multiple response question; in most cases they mentioned only one or two items from the code list.

*3.2.2 Variables in **Wrq98.**:*

Variables arising from the Worker Representative questionnaire have a two-character prefix. The first character (W) is short-hand for Worker Representative. The second character signifies the section of the questionnaire from which the variable arises. So **WAREPTYP** arises from Section A of the Worker Representative questionnaire. Variables arising from multiple-response questions are labelled in the same way as in **Mq98fin.***

*3.2.3 Variables in **Seq98.**:*

A one-character prefix points to the relevant section of the Survey of Employees questionnaire. Questions inviting more than one box to be ticked (**B1**, **B3** and **D3**) yield one dichotomous variable for each of the possible responses (i.e. **B11** to **B15**). An additional variable with the same name as the question (**B1** in this example) indicates the number of boxes ticked by the respondent.

Note: **A6** was not intended to elicit multiple responses but was multi-coded by a number of respondents. Hence, there are two versions of the variable: first, a single-coded variable named **A6** which takes the value of 0 if more than one box was ticked; and second, a multiple-response variable **A6MULT** which takes the form outlined in the previous paragraph.

*3.2.4 Variables in **Pq_9098.*** & **Pq_98out.**:*

The panel data files incorporate data from both the 1990 Cross-Section survey and the 1998 Panel survey.

Variables originating in Management data file of the 1990 Cross-Section have a single-letter prefix that identifies a particular section of the 1990 Main Management questionnaire, from A to L. The remainder of the variable name then usually consists of a number, relating to the question number within that section (e.g. **A14**). The exceptions are variables originating from the 1990 Basic Workforce Data Sheet, which use more descriptive variable names (e.g. **TOTEMP**, **MANFTM**).

Variables from the 1990 questionnaire for Worker Representatives of Manual Employees are prefixed with the letters MA to MK. Those from the 1990 questionnaire for Worker Representatives of Non-Manual Employees are prefixed with the letters NA to NK. Variables prefixed FA through to FC and contain data from 1990 interviews with Financial Managers.

Panel data collected in 1998 is contained within variables that are prefixed with the letter Y. This prefix is followed by a second letter indicating the relevant 1998 questionnaire section. So, the variable **YBSTATUS** arises from Section B of the 1998 Panel questionnaire. The letter X is used to prefix derived variables from 1990 (e.g. **XBSTATUS**). The remainder of the variable name is taken from the equivalent variable in the 1998 Panel questionnaire. So, **XBSTATUS** (the derived variable

indicating the formal status of the establishment in 1990) is so named because it is derived to be equivalent to **YBSTATUS** in 1998, although it originates from the 1990 variable **A3**.

Note: multiple-response items use different naming conventions within the 1990 and 1998 data in the Panel data file. Variables arising from multiple-response items within the 1990 Cross-Section, have a suffix of the form _d1, _d2 etc. (e.g. **B18_d1**, **B18_d2** and so on). Here, the number refers to the order of the response on the code frame. The _d indicates that the variables is dichotomous, with each variable containing a 1 if that particular response was mentioned in the interview. So, **B18_d2** contains a 1 if the second code on the code frame for **B18** ('Management consultant') was mentioned. Otherwise, the variable contains a zero, unless the respondent did not answer **B18** at all, in which case it will be missing. Variables arising from multiple-response questions in the 1998 panel interview are numbered with the order of the response, as in **Mq98fin.***. However, the number is preceded by an underscore (as in the case of **YPCOM_1** to **YPCOM_8**).

3.3 The layout of the data files

*3.3.1 **Mq98fin.*** and **Wrq98.**:*

In both **Mq98fin.*** and **Wrq98.*** the first variable is **SERNO**. This is the unique workplace identifier.

The unique workplace identifier enables the user to match data together from different files. For example, one can combine information from **Mq98fin.*** with that from **Wrq98.*** in order to compare managers' and worker reps' reports of union membership density at the workplace. Alternatively, one might combine information from **Mq98fin.*** with that from **Seq98.*** in order to assess the degree to which employees' attitudes vary by industry or size of workplace. The process of matching of data from different data files using SPSS or STATA is outlined in Chapter 6 of this Guide.

Following the unique workplace and employee identifiers, the next variables to appear in **Mq98fin.*** and **Wrq98.*** are the weight variables. These are outlined in more detail in Chapter 4. Then follows a set of variables labeled **XCODE1** to **XCODE5**, and **ZALLEMPS**. The **XCODE** variables are used to indicate cases that have been edited in some particular way by the research team, or cases for which questions still remain about the validity of some aspect of the data. Further details are provided in Section 6.7 of the WERS98 Technical Report (Airey *et al*., 1999). **ZALLEMPS** gives the number of employees employed at the establishment at the time of interview.

The remaining variables in **Mq98fin.*** and **Wrq98.*** follow in the same order as they appear in the relevant questionnaire. The variable names are replicated from the questionnaire document.[2] Note, however, that **Mq98fin.*** contains a full set of Z-prefixed variables from the Employee Profile Questionnaire; only those generated by computer calculation within the interview are actually listed in Part 1 of the

[2] Departures from this rule are cited in the volume of Variable Notes mentioned in Section 2.1 of this Guide.

Management Questionnaire document. Both files end with a very small selection of variables derived by the WERS98 research team during primary analysis. These derived variables are prefixed with the letter N.

*3.3.2 **Seq98.**:*

The variables on the data file from the Survey of Employees, **Seq98.***, follow much the same pattern, with one important exception. In this file the first variable is the unique *employee* identifier **SERIAL**. The *workplace* identifier (**SERNO**) is the second variable on this file. After **SERNO**, the variables follow the order of the questions in the Survey of Employees questionnaire.

The final variable on **Seq98.*** is the weight, **EMPWT_NR**. The issue of weighting is discussed below in Chapter 4 of this Guide.

*3.3.3 **Pq_9098.*** & **Pq_98out.**:*

In both panel data files, the data from the 1990 Cross-Section precede those arising from the 1998 Panel survey.[3]

In both files, the 1990 variables begin with **SERNO2**, the unique workplace identifier.[4] Variables then follow in accordance with the order of questions in the Main Management questionnaire from the 1990 Cross-Section survey. Data from the 1990 Basic Workforce Data Sheet are followed by data from Sections A, B and so on through to Section P. Then follows data from interviews with worker representatives of manual employees (variables prefixed by the letter M) and worker representatives of non-manual employees (prefixed N), where present. The final group of 1990 variables (prefixed F) contain data from interviews with Financial Managers, where present. A single derived variable, **XBSIC80B**, is located at the end of this group.

Note: The WERS98 User Guide does not incorporate documentation on the 1990 Cross-Section survey. This documentation may be obtained separately from the Data Archive: see Appendix B.

In **Pq_9098.***, the 1990 variables are followed by variables containing data from the 1998 panel survey interviews. These begin with a variable **EDITOUT** that contains an outcome code for each interview. Variables then generally follow their order in the 1998 panel questionnaire, from **YAALLEMP** to **YVURELS**. The section of variables with the prefix YZ contains administrative data concerning the interview.

---

[3] STATA users should note that their versions of the two WERS98 Panel Survey data files on general-release have each been divided up into two or three components, so as to comply with STATA's limitations on the maximum number of variables permitted within a single file. A 'ReadMe' text file, sent with the data files by the Data Archive, explains the division of the data between the files.

[4] Users should note that this is a new variable and does not match the serial number on the original 1990 cross-section data file, deposited in 1992. For reasons of confidentiality, the 1990 variables giving the workplace's detailed industry classification and regional location have been moved into a restricted-access data file. The original 1990 serial number has been changed to inhibit users from simply matching this data back on from a copy of the full 1990 file.

Following those variables with the YZ prefix comes a set of variables, from **YBEMD_1** to **YVISYR_5**, which contain numeric codes that have been derived from the answers to open-ended questions in the 1998 panel questionnaire.[5] Then follows a variable **YZOVCOD1** which is used to indicate cases that have either been edited in some particular way by the research team or for which questions still remain about the validity of some aspect of the data (see Section 6.7 of the WERS98 Technical Report (Airey *et al.*, 1999)). **YZLOC** is merely a replica of **EDITOUT** – see above.

A set of variables prefixed by the letter X follow **YZLOC**. These variables – **XZMONTH** to **XFUNR14** - contain data from the 1990 cross-section interview that was fed forward into the 1998 panel interview for the purposes of identifying change. Then follow **YZYEAR** and **YZMONTH** giving, respectively, the year and month of the 1998 interview. After **YZMONTH**, the remainder of the data file consists of those elements of the 1998 BWDS not punched during the interview (**YAUSKFTM** to **YAMGRPTF**); a handful of derived variables (**YEUDENS** and **YBSIC80B**); a small number of questionnaire variables that were relocated during the preparation of the file (**YG90CHK1** to **YVFINBLW**); and, finally, the weight variable **PWEIGHT**.

The second panel data file, **Pq_98out.***, has a much simpler layout. This file consists of data from the 1990 cross-section survey, as described above, and one additional variable, **EDITOUT**, which contains a 1998 outcome code for each workplace that yielded a productive interview in the 1990 Cross-Section survey.

3.3.4 Restricted data files:

The restricted data files are of two types: data files and Excel spreadsheets. Details of the restricted files are given in Tables 1 and 2 of Appendix A.

Each of the data files begins with the unique workplace identifier (**SERNO** or **SERNO2**), after which follow the restricted data items.

The Excel files of verbatim responses from the Management, Worker Representative and Panel interviews (**Mqopen.xls**, **Wrqopen.xls** and **Pqopen.xls**) contain one sheet per question. On a particular sheet, each row contains a unique workplace identifier (**SERNO**), the numeric code to which the verbatim was assigned and the verbatim response itself, as given by the respondent in that workplace.

The Excel file relating to the Survey of Employees (**Seqopen.xls**) contains verbatim text from a single question, **D12**. The verbatims span several sheets and are arranged in batches relating to the time of their arrival in the fieldwork office. Each row contains the unique employee identifier (**SERIAL**) and the text written at **D12** by that respondent.

Note that the answers contained in all four of the Excel spreadsheets have been anonymized in order to protect the confidentiality of respondents. This means that all references to organization names or individuals have been replaced by a string of

[5] In some cases, variables numbered _2 or above are devoid of data (e.g. **YBSIC_3** to **YBSIC_5**; **YBEMI_4** and **YBEMI_5**). This indicates that all respondents gave fewer than the maximum number of responses allowed in the interview (generally 5).

xxxxx's. Further information on the Excel spreadsheets of verbatim answers is given in Section 6.4 of this Guide.

3.3.5 Final note:

The user should be aware that there are a number of questions from the Management, Worker Representative and Panel questionnaires which do not have corresponding variables in the deposited data files. These questions, which generally collected confidential information such as the name of the establishment or the organization to which it belonged, have been dropped in order to preserve the anonymity of respondents. Such questions are clearly marked in those versions of the questionnaires that are available from the Data Archive or the WERS98 Data Dissemination Service web site. They are also listed in the volume of *Variable Notes* produced by the Data Dissemination Service.

4. Weighting

Weighting is crucial to the analysis of WERS98. However, it is also an issue that creates much confusion. The aim of this section is to explain both the principle and practice of weighting in respect of WERS98, so that the issue is both better understood and more confidently addressed by users.

There are two key rules to follow in respect of weighting:

1. Weighting must be applied to all analyses to account for the WERS98 sample design, if one is to obtain unbiased population estimates from the survey data.
2. One must also account for the features of the WERS98 sample design in the calculation of standard errors and the application of significance tests, if one is to obtain accurate estimates of the reliability (precision) of the survey data.

The rationale behind these two rules is set out in various sections below. The principles of weighting each of the WERS98 data sets are first outlined. Users are then given the names of the various weight variables present in WERS98, and instructions on how they can be applied and removed in SPSS and STATA. Finally, the implications for statistical inference are explained, with instructions being given as to how standard tests can be adjusted for use with WERS98 data.

4.1 The 1998 Cross-Section data (Managers and Worker Reps)

4.1.1 Principles of weighting the 1998 data from managers and worker reps

Each of the cross-section surveys in the WIRS series is a sample survey, meaning that interviews are undertaken with only a selection (or sample) of eligible workplaces within the population.[6] As long as the process of selecting the issued sample (the sample distributed to interviewers) is essentially random, and the rate of response to the survey does not differ to any substantial degree between different types of workplace, those workplaces that eventually take part in the survey (the achieved sample) will constitute an unbiased, representative sample of all workplaces in the population from which they have been selected. Results from these workplaces can then be generalized to the population as a whole.

The sampling procedure used in WERS98 is outlined in some detail in the Technical Report (Airey *et al*., 1999). The most pertinent point to note for the purposes of this section on weighting, however, is that the issued sample of workplaces was arrived at through a process of *stratified random sampling* using *variable sampling fractions*.

The population of workplaces in Britain is dominated by small workplaces, and comprises many more workplaces in manufacturing than it does in construction for example. A process of *simple random sampling* from this population would therefore generate a similarly distributed sample which, unless it contained a very large number of units overall, would not include sufficient large workplaces (or construction

[6] The alternative would be to take a census, whereby all eligible workplaces in the population would be surveyed.

workplaces) to permit reliable inferences to be drawn for such groups. The use of stratification and variable sampling fractions overcomes this problem whilst retaining the necessary element of random selection. The population is first divided (or stratified) into distinct groups (or strata). A separate random sample is then taken within each stratum, using sampling fractions that vary according to the particular stratum. The process of stratification ensures that one selects the correct number of cases from within each stratum of the population, whilst the use of variable sampling fractions enables one to select sufficient cases to be able to analyse each stratum separately.

In the case of the WERS98 cross-section, the population of workplaces recorded on the sampling frame – the Interdepartmental Business Register (IDBR) - was stratified using six categories of workforce size and twelve Major Groups (D to O) of the 1992 *Standard Industrial Classification*. A unique sampling fraction was then applied to each of the 72 resultant strata. Sampling fractions increased with employment size, whilst units were over-sampled in Major Groups E, F, H, J and O, and under-sampled in Major Group D. This design ensured that, within the overall selected sample of 3192 units, there were at least 100 units in each Major Group and at least 350 units in each of the six workforce-size categories.

The 1998 cross-section survey achieved a very high response rate (80 per cent), which did not vary to any substantial degree by either workforce size or industrial classification. Hence, the achieved sample retained a very similar profile to that of the sample initially selected from the IDBR. However, the use of variable sampling fractions means that the profile of the achieved sample (or the initial sample) did not match that of the population from which it had been derived.

The sample must therefore be adjusted in order to eliminate this distortion before unbiased estimates can be derived about the population that the sample is intended to represent. Failure to do so can lead to seriously misleading results. The distortion is eliminated by attaching *differential sampling weights* to the sampled units prior to analysis. For any one unit, this weight is equal to the inverse of that unit's probability of selection into the sample. If the probability of selection of a particular unit is ¼, the value of the weight will be 4. This single unit will then represent 4 units in any weighted analysis.

In most cases, the probability of selection of a particular workplace within the WERS98 Cross-Section could simply be taken as the sample fraction imposed on the sample stratum from which it originated. However, in some cases, adjustments had to be made to this sample fraction in order to arrive at a more accurate estimate of the true probability of selection. Extreme weights were also trimmed. (See Section 7.1.1 of the WERS98 Technical Report for further details.)

4.1.2 Weight variables to be used in analysis of 1998 data from Managers and Worker Reps

There are two variables that can be used to weight the WERS98 data from Managers and Worker Reps. These are **EST_WT** and **EMP_WT**.[7] The first of these, **EST_WT**, is used for workplace-level analysis, whilst the second, **EMP_WT**, can be used to generate employee shares (see below).

EST_WT is the standard establishment-level weight, representing the inverse of the probability of selection of each establishment into the survey sample (notwithstanding the trimming of extreme weights, mentioned in the previous section). Each weight was divided by a scaling factor (approximately 117) during the derivation of **EST_WT** so that the total weighted number of workplaces sums to 2191: the number of cases in the achieved sample. **EST_WT** has a range from 0.01 to 10.24, with around 90 per cent of cases having values below 2.20.

EMP_WT can be used to produce analyses which reflect the proportion of employees (not workplaces) to whom a particular workplace characteristic pertains. It has been derived by multiplying the workplace weight (**EST_WT)** by the total number of employees at the workplace at the time of interview (**ZALLEMPS**), then dividing this product by a scaling factor which brings the overall weighted base back to 2191: the number of cases in the achieved sample. The scaling factor is equal to the average number of employees found in workplaces in the sample (approximately 62). **EMP_WT** has a range from 0.05 to 31.08, with around 90 per cent of cases having values below 1.80.

4.1.3 A practical example of the difference between weighting schemes

The different uses of the two weights can be seen by separately analysing one item of data under both weighting schemes. Take the variable **CATESTS**, which indicates whether a workplace uses personality or aptitude tests when filling vacancies.

Using unweighted data from **Mq98fin.***, we see that 33 per cent of all workplaces in our sample use personality or aptitude tests when filling vacancies. However, further investigation shows that this practice is more common amongst larger establishments. Larger establishments are over-represented in our unweighted sample when compared with the population as a whole, because of the sample design, and so we can expect the use of personality or aptitude tests to actually be lower when we look beyond our sample to the population at large.

This is confirmed by applying the workplace weight, **EST_WT**, which restores the profile of the sample to that of the population. Under this weighting schema, we arrive at a population estimate of 19 per cent.

But what about the proportion of *employees* that work in such workplaces? Since larger workplaces are more likely to use personality or aptitude tests, we can expect

[7] A third weight variable, **GROSSWT**, is present on the Management and Worker Representative data files but should not be used as it is now thought not to provide accurate gross numbers of workplaces. A fourth weight, **EST_WT1**, is present on the Worker Representative data file only – this is equivalent in function to **EST_WT** and can also be ignored.

the proportion of employees working in establishments where personality or aptitude tests are used to screen applicants to be greater than 19 per cent. Analysis of **CATESTS** using **EMP_WT** provides an estimate of 36 per cent.

4.2 The 1998 cross-section data (employees)

4.2.1 Principles of weighting the 1998 data from employees

The Survey of Employees was based on a two-stage sample design. The selection of workplaces into the sample for the Main Management interview represented the first stage; the selection of employees within each of those workplaces represented the second stage. Readers are therefore advised to have read Section 4.1 above before proceeding.

Within each workplace taking part in the WERS98 Cross-Section, a sample of 25 employees were selected to participate in the Survey of Employees. In workplaces with between 10 and 24 employees, all employees were asked to participate. These 25 (or fewer) employees were selected at random from a list of all those employed at the workplace; the selection procedure is outlined in the WERS98 Interviewer Training Manual (Volume 7 in the WERS98 User Guide).

The resultant data can be analysed in two ways, both requiring a different system of weighting.

First, the data can be analysed independently as a survey of all employees working within workplaces that have 10 or more employees in total. In order to derive unbiased estimates about this population from the survey data, the data must be weighted to take account of the probability of selection of each employee into the sample. This probability is derived as the multiple of:

a) The probability of selection of the employee's workplace into the sample of workplaces, and
b) The employee's own probability of selection from among the employees at that workplace.

The weight is then calculated as the inverse of this probability.

The rationale for taking account of the probability of selection of each workplace is set out in the previous section. The employee's own probability of selection within each workplace also needs to be taken into account since the use of a fixed sample size within workplaces of 25+ employees meant that the overall proportion of employees from very large establishments that were asked to complete a questionnaire was much lower than the overall proportion asked from establishments with smaller workforces. Employees from small establishments would therefore be over-represented in the final achieved sample of employees if such an adjustment was not made.

The previous section stated that there was no apparent response bias among the achieved samples of Managers and Worker Reps. However, an analysis of response to the Survey of Employees found that certain groups of employees (e.g. part-time workers) were less likely to return their questionnaire than others. This meant that,

even after taking account of differing selection probabilities, certain groups were still either under or over-represented in the final achieved sample when compared with the population as a whole. The weights therefore needed to be adjusted in order to remove any bias that may have been introduced by employee non-response. Further details may be found in Sections 7.1.4 and 7.1.5 of the WERS98 Technical Report. The final employee weights produced by these various stages are found in the standard Survey of Employees weight (**EMPWT_NR**).

The second way in which the Survey of Employees data can be analysed is at workplace-level. Here, the data collected from each employee is combined with that collected from other employees in the same workplace to produce summary information about the workforce as a whole within that establishment. For example, one might use the returned employee questionnaire data to compile a measure of the average level of satisfaction among employees at that establishment. The process of combining employee records to produce summary measures at workplace level is described in Section 6.3.

Since the selection of employees within each workplace is random, one does not have to address the issue of variable sampling fractions between employees in same workplace, although one should compile aggregated measures from weighted data, so as to account for non-response bias. Then, when the data is aggregated to workplace level, one must take account of workplace-level sampling by applying the workplace-level weight (**EST_WT**).

In this second type of analysis, the fixed sample size could lead to concerns about the generalizability of the data collected in large workplaces. In essence, one must be confident that one has enough employee returns to be able to summarise the variation present among the workforce at a particular establishment. This issue is dealt with in more detail in Section 6.4 of this Guide.

4.2.2 Weight variable to be used in analysis of 1998 data from employees

When the Survey of Employees data is to be analysed with the employee as the unit of analysis (the first mode of analysis described above), the weight variable that should be used is **EMPWT_NR**. This is the only weighting variable that is available on the Survey of Employees data file. With **EMPWT_NR**, the weighted number of employees sums to 28,222: just slightly more than the number of cases in the achieved sample (28,215). **EMPWT_NR** has a range from 0.04 to 17.82.

When the data is to be analysed with the workplace as the unit of analysis (the second mode described above), the workplace-level weight **EST_WT** should be used.

4.3 The 1990-98 panel data (PQ_9098.)*

The 1990-98 panel data consists of two observations. The first derives from the Management interview in the 1990 WIRS Cross-Section Survey, the second from the WERS98 Panel Survey.

4.3.1 Principles of weighting the 1990-98 panel data

Given the nature of the panel data, account needed to be taken of two rounds of sample selection and potential non-response in compiling a weight.

As stated above, the first wave of the 1990-98 Panel (the 1990 observation) is provided by the Management interview from the 1990 WIRS Cross-Section survey. The sample design used in the WIRS90 Cross-Section was similar to that used in the WERS98 Cross-Section (except that there was a much smaller degree of differential sampling by industry) and so the weight for the 1990 Cross-Section was derived in broadly same way as outlined in Section 4.1.

The initial sample for the second wave of the Panel (the 1998 observation) was taken as a 63% (1301/2061) random sample of productive workplaces from the 1990 Cross-Section. This sampling fraction of 63% was applied equally within 7 strata, defined according to workforce size in 1990. As the sampling fraction was equal within each stratum, productive cases from the 1998 wave of the Panel Survey would then be a representative sample of productive workplaces from the 1990 WIRS Cross-Section that were still in existence and in-scope in 1998. That is, as long as there was no response bias.

Analysis showed that there was some bias in the level of response between different types of workplace in 1998, with certain parts of the public sector being more likely to respond, for example. This meant that the productive cases from the 1998 wave were not fully representative of the initial 63% sample. The final panel weight therefore needed to incorporate an adjustment for non-response bias.

Putting these elements together, the final sample of productive interviews from the 1998 wave of the Panel Survey can be made to represent the initial sample of productive cases from WIRS90 that were still in existence and in-scope in 1998 by applying the inverse of the sampling fraction (2061/1301), together with an adjustment for non-response. The WIRS90 weight is then applied in order to adjust for the stratification of the WIRS90 sample on which the Panel Survey was based.

4.3.2 Weight variable to be used in analysis of the 1990-98 panel data

A single weight, **PWEIGHT**, incorporates each of the elements of weighting outlined above. This weight is used irrespective of the wave from which the variable of interest derives. In other words, **PWEIGHT** is used whether one wishes to analyse the incidence of joint consultative committees in 1990 (**XPJCC**) or 1998 (**YPJCC**). When **PWEIGHT** is applied, the total weighted number of workplaces sums to 881. **PWEIGHT** has a range from 0.01 to 5.39

4.4 The 1998 outcomes data (Pq_98out.*)

The 1998 outcomes data file consists of a single 1998 outcome code (e.g. closed down, survived etc.), which has been matched onto the data obtained in the 1990 WIRS Cross-Section survey. A 1998 outcome was identified for each of the 2,061 productive cases in the 1990 Cross-Section survey.

4.4.1 Principles of weighting the 1998 outcomes data

Since there are no new sampling issues to address, the weighting for the 1998 outcomes data file is simply that pertaining to the 1990 Cross-Section. As stated in the previous section, the sample design used in the WIRS90 Cross-Section was similar to that used in the WERS98 Cross-Section, except that there was a much smaller degree of differential sampling by industry, and so the weight for the 1990 Cross-Section was derived in broadly same way as outlined in Section 4.1.

4.4.2 Weight variable to be used in analysis of the 1998 outcomes data

The variable named **WEIGHT** is used to weight the 1998 outcomes data. Other weight variables present on the data file (**WEIGHT1** and **WT2**) can be ignored. When **WEIGHT** is applied, the total weighted number of workplaces sums to 2,000. **WEIGHT** has a range from 0.01 to 4.37.

4.5 Applying and removing weights

4.5.1 Applying and removing weights within SPSS

Users should note that some of the SPSS data files come with the weight already applied to the data (see Section 3.1). In other words, there is no need to apply the weight yourself before you begin to analyse data in these files. Other data files are unweighted when you load them into SPSS, so you need to apply the weight before you can get unbiased population estimates from the data.

Users should also note that some of the standard procedures in SPSS, such as `crosstabs`, do not adequately deal with the non-integer weights that are a feature of WERS98. Specifically, `crosstabs` will round the weighted counts in each cell to integers before calculating column or row percentages. This can generate misleading results, particularly when the weighted counts are small. The SPSS Tables module, described in Chapter 5 of this Guide, does not have the same problem. This is one of the reasons why we would consider SPSS Tables to be preferable for conducting tabular analysis of WERS98.

To apply the weight **EST_WT** in SPSS:

(i) Using syntax, type:

```
weight by est_wt .
```

(ii) Using the menu system:

In the Data Editor, select ‘Weight cases’ from the drop-down menu headed ‘Data’. Highlight **EST_WT** from the list of variables. Check the ‘Weight cases by’ radio button and click on the arrow to transfer **EST_WT** into the box headed ‘Frequency variable’. Click on ‘OK’.

Whether using syntax or menus, when the weight has been applied the phrase 'Weight on' will appear in bottom row of Data Editor, towards right hand side of screen. All subsequent analyses will be run on weighted data until the weighting is removed or the data file closed.

To remove weighting in SPSS:

(i) Using syntax, type:

```
weight off .
```

(ii) Using the menu system:

In the Data Editor, select 'Weight cases' from the drop-down menu headed 'Data'. Check the 'Do not weight cases' radio button. Click on 'OK'.

The phrase 'Weight on' will disappear from bottom row of Data Editor. All subsequent analyses will be run on unweighted data.

4.5.2 Applying and removing weights within STATA

STATA recognises a number of different types of weight variable (see Section 14.1.6 in the STATA User Guide). The weights used in the analysis of WERS98 are what STATA refers to as sampling weights, or pweights. Here, pweights refers to 'probability weights' and is not to be confused with the 1990-98 Panel Survey weight variable **PWEIGHT**. Sampling (or probability) weights can be handled in two different ways within STATA:

(i) Using the `svy` family of commands:

The `svy` family of commands within STATA have been specifically created for the analysis of data arising from complex survey designs. This means that, through the svy commands, one can not only apply a weight, but also ask STATA to take account of the sample design when calculating standard errors. Specifically, the `svy` commands can take account of both the probability sampling and the stratification that featured in the design of the WERS98 workplace samples. They can also take account of the clustering of employees within workplaces when analysing the Survey of Employees. An overview of the commands is given in Chapter 30 of the STATA User Guide. An explanation of why this is important is contained in Section 4.6.

There are four pieces of information about the WERS98 sample design that STATA can use with its `svy` commands. These are:

1. The final weight
2. The nature of the sample stratification
3. The sampling fractions used to select workplaces in each stratum
4. The clustering of employees within workplaces.

Weights and sample strata (items 1 and 2) should be specified at all times, whether analysing data from the WERS98 Cross-Section or the Survey of Employees.

Sampling fractions (item 3) should only be specified when conducting workplace-level analysis. Sampling fractions should not be specified when analysing data from the Survey of Employees because of the multi-stage nature of the survey design - see Section 30.2.2 of the STATA User Guide. The clustering of the employee sample (item 4) should naturally only be specified when conducting employee-level analysis.

In respect of the Management data from the WERS98 Cross-Section, the weight (**EST_WT**) is available from the data file on general release. Items 2 and 3 are not. However, the way in which the sample frame was stratified prior to selection, and the sampling fractions used, are reproduced in Tables 2A and 2B of the WERS98 Technical Report. A file that specifies the stratum from which each productive workplace originated, along with the relevant sampling fraction, has been created by the WERS98 Data Dissemination Service and is available from the Data Archive (filename: **Sample98.***). The strata are identified in the variable **IDBRSTR2**, whilst the sampling fractions are contained in **IDBRSF2**. So having read in the WERS98 Cross-Section Management data file, the `svyset` command would be used to inform STATA about the design of the WERS98 workplace sample in the following way:

```
svyset pweight est_wt
svyset strata idbrstr2
svyset fpc idbrsf2
```

In respect of the Employee data from the WERS98 Cross-Section, the weight (**EMPWT_NR**) is available from the file on general release. The strata are available in the file **Sample98.***, as mentioned above. The clusters are specified using the workplace identifier (**SERNO**), which is part of the general release file. So having read in the WERS98 Cross-Section Employee data file, the `svyset` command would be used to inform STATA about the design of the WERS98 employee sample in the following way:

```
svyset pweight est_wt
svyset strata idbrstr2
svyset psu serno
```

Having told STATA about the sample design and weighting, one can then begin to use the descriptive and analytic commands in the `svy` family (e.g. `svytab`, `svymean` and `svyreg`). More is written about STATA's `svy` commands in Chapter 30 of the STATA User Guide.

Users should note that the sample data provided in variable **IDBRSTR2** in **Sample98.sav** will enable you to make adjustments for sample stratification and sampling fractions when running analyses of the full WERS98 management data file of 2191 cases. However, users will encounter problems in the analysis of sub-samples (e.g. private sector) or of variables with many missing values. This is because STATA will not run `svy` commands on sub-samples in which there is only one observation in a particular sample stratum.[8] Users can easily get around this restriction by grouping strata on **IDBRSTR2** until new groups are formed that contain

[8] This is the reason why the variable **IDBRSTR2** has only 71 categories, compared with the 72 on the original sample stratification variable **IDBRSTR1**.

more than one observation (see entry for `svydes` in the STATA Reference Manual). This new, grouped variable can then be specified at the strata option on `svyset`.

When grouping two strata together, it is advisable to collapse ones that account for a similar number of units in the population (see Table 2A of the WERS98 Technical Report) and that can be expected to have similar population values for items covered by WERS98. An advisable initial strategy, therefore, is to collapse strata representing adjacent size categories within the same SIC92 Major Group.

It is much more time-consuming to calculate sampling fractions for the new strata. This can be done by using the information in Tables 2A and 2B in the WERS98 Technical Report. However, specifying the sampling fractions using `fpc` reduces the standard errors and so omitting to tell STATA about them is equivalent to adopting a conservative approach in the evaluation of statistical significance.

(ii) Specifying pweights in non-`svy` commands:

Weighted analysis can also be produced by specifying the weight variable as a sampling weight, or pweight, within the options available on most of STATA's non-`svy` commands. For example:

```
xi: regress eunionum ztu_mem i.astatus [pweight=est_wt]
```

Note, here, that the use of pweights with STATA's non-`svy` commands will generate the same point estimates as produced by the equivalent `svy` command. However, standard errors will be slightly less accurate under the non-`svy` approach. See Section 23.13.3 and Chapter 30 the STATA User Guide for more details.

Note also that some of the common non-`svy` commands that produce descriptive statistics, such as `tabulate` and `summarize`, do not permit the specification of pweights. `Svytab` and `svymean` are the relevant alternatives from the `svy` family. Specifying an aweight, rather than a pweight, on `tabulate` or `summarize` will generate the correct point estimates (cell proportions in the case of `tabulate`; means in the case of `summarize`). However, `tabulate`'s weighted cell counts are not accurate (they are scaled by a factor equal to: [Unweighted base for table/ Weighted base for table]). For its part, `summarize` displays the standard deviation of the sample observations, whilst `svymean` displays the standard error of the estimated population mean.

4.6 The implications of sample design for statistical inference

It has already been established in Sections 4.1 to 4.4 that the design of the WERS98 sample has the effect of introducing *bias* to any estimates that are derived from the raw data. As a result, one must account for the sample design by applying weights to the data, if one wishes to obtain unbiased population estimates. However, the sample design also affects the *reliability* of the estimates from WERS98. Put simply, if we do not take account of the sample design, we are likely to overstate the reliability (or precision) of our estimates.

All calculations that are derived from samples have a degree of sampling error. In other words, even after we have removed any bias, our sample can still only provide us with an estimate of the true population value, and this estimate naturally has some degree of imprecision (called sampling error). The degree of sampling error depends upon three factors: the degree of variability in the population; the size of our sample (and in extreme cases sampling fraction); and the way in which the sample has been constructed (Hedges, 1978: 60). In broad terms, the sampling error increases with the degree of variability in the population, decreases with sample size and increases with the complexity of the sample.

Fortunately, sampling errors can be estimated through standard formulas, enabling us to formally assess the reliability of our sample estimates. This point can be illustrated by referring to the standard formula for estimating the sampling error of a sample mean (e.g. the mean number of union members in a workplace).

The standard formula is as follows:

$$s.e.(\bar{x}) = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n^2} \cdot \left(1 - \frac{n}{N}\right)}$$

where: $\bar{x}$ is the sample mean, *n* is the number of observations in the sample and N is the number of cases in the population (such that $\frac{n}{N}$ represents the sampling fraction). The last term $(1 - \frac{n}{N})$ is called a 'finite population correction' and is generally omitted unless the sampling fractions are greater than 0.10. It is included here for completeness.

This statistic gives you what is called the 'standard error' of the sample mean. Statistical theory says that we can be 95 per cent confident that the true population value lies within an interval of two standard errors either side of our sample value. Different formulas exist for calculating the sampling errors associated with: proportions (percentages); differences between means or between proportions; regression coefficients; and degrees of dependence (or independence) between variables. See Sections 4.6.1 to 4.6.3 for further details.

The example clearly illustrates that the standard error is determined by the variability present in the sample $\left(x_i - \bar{x}\right)$; the sample size *n*; and, in extreme cases (e.g. sampling fractions greater than 0.10), the sampling fraction. Specifically, we can see that the degree of reliability (or precision) in our sample estimate will be greater if the values in the sample are less dispersed, if our sample size is greater and (in extreme cases) if our sampling fraction is large. The influence of the sample size shows why it is important to consider the unweighted number of cases on which any sample estimate is based.

The one thing that this formula does not account for, however, is the sample design. This is because the normal procedures for calculating standard errors (whether of means, proportions, differences between proportions or in multivariate analysis) and the standard means of assessing significance or independence all assume that the estimate has been derived from a basic sample design. This basic sample design is called 'simple random sample with replacement' (SRSWR). SRSWR means that the sample is formed by simply of taking a random selection of cases from the population, using a fixed sampling fraction for all cases, and using a method whereby each case is available for re-selection, even if it has already been sampled (hence the term 'with replacement').

Unfortunately, WERS98 was not based on a SRSWR design, but a more complex sample design that gives larger sampling errors. Specifically, the workplace sample for the WERS98 Cross-Section was derived by applying unequal sampling fractions with different strata of the population, whilst the Employee sample also incorporates clustering (since employees are only sampled if their workplaces have already been selected for the workplace sample). The effect, in both cases, is to increase sampling errors when compared with SRSWR designs. Standard methods of estimating the sampling error associated with estimates from the survey are therefore no longer valid and will give misleading results, leading us to conclude that the WERS98 estimates are more reliable (precise) than they really are. Hence, we need to adjust the standard methods of estimating the sampling error in order to account for the more complex sample design used in WERS98.

A statistic called the 'design factor' (*deft*) gives a measure of the degree of amplification in sampling errors that results from using a complex sample design rather than SRSWR (Kish 1965). So, if we know the *deft* associated with a particular estimate, we can use it to correct the standard formula and estimate the true sampling error under the complex sample design.

The design factor associated with a particular estimate (e.g. a mean or proportion) is calculated as the ratio of its standard error under the complex design to the standard error that would apply in a SRSWR of the same (unweighted) sample size.[9] Formally:

$$s.e.(\bar{x})_{COMPLEX} = deft \cdot s.e.(\bar{x})_{SRSWR}$$

The *deft* for individual estimates can be calculated in STATA by using the `svy` family of commands. This is not possible in SPSS, but the *deft* has already been calculated for a wide range of variables from the WERS98 Cross-Section and Panel Surveys. These *defts* can be found in the tables in Section 8.1 of the WERS98 Technical Report. If the variable you are analysing does not feature in these tables, its *deft* can be most closely approximated by using the *deft* for a variable with which it is closely correlated. A less accurate alternative is to use the average *deft* that has been calculated for each survey.

The WERS98 Cross-Section Main Management survey is estimated to have an average design factor of 1.5 (Airey *et al.*, 1999: 95). This means that the standard

[9] The square of the design factor is called the 'design effect' and is the ratio of the two variances (since the variance is the square of the standard error).

errors associated with particular estimates from the Main Management interview are, on average, 1.5 time larger than they would have been had the survey been conducted under SRSWR. The Survey of Employees is estimated to have an average design factor of 1.7 (Airey *et al.*, 1999: 104)

As a result, if one merely uses the standard formulas for calculating sampling errors and the normal tests of statistical significance or independence, each of which assume SRSWR, one could make many Type I errors since you are assuming that the sample is more reliable (precise) than it is in practice. The various ways to adjust the standard formulas and tests are further outlined below.

4.6.1 Frequency analysis

By frequency analysis we mean analysis that aims to estimate either:

a) the proportion of the population (or a sub-population) with a particular characteristic, or
b) the mean value of a particular variable in the population (or a sub-population).

First, consider (a):

Taking a real example from WERS98, running a weighted frequency of **IPOLICY** on private sector workplaces (**ASTATUS**<3) tells us that 57.3 per cent of all private sector workplaces had a formal written policy on equal opportunities. This is based on an (unweighted) sample size of 1507. We wish to know how reliable this estimate is (in other words, what it enables us to say about the population).

The formula for the standard error of a proportion under SRSWR is as follows:

$$s.e.(p) = \sqrt{\frac{p(1-p)}{n}}$$

where p is the proportion in question. We have ignored the finite population correction term in this formula, for simplicity.

The SRSWR standard error of our proportion is therefore 1.3. So, under SRSWR, we could be 95 per cent confident that the proportion of private sector workplaces *in the whole population* that have a written policy on equal opportunities lies between 56.0 per cent and 58.6 per cent (or between 56 per cent and 59 per cent, after rounding).

However, Table 8A of the WERS98 Technical Report shows that **IPOLICY** has a design factor of 1.9. The true standard error of **IPOLICY** under the WERS98 sample design is therefore 1.9*1.3=2.5 (after rounding). Accordingly, we can actually only be 95 per cent confident that the true population value lies between 55 per cent and 60 per cent (after rounding). The true confidence interval is therefore almost double that suggested by the uncorrected formula (5 per cent, compared with 3 per cent). This is the true measure of the reliability (precision) of our estimate of 57 per cent.

Considering (b):

Again, we take a real example from WERS98. A weighted mean of union density (using a derived variable that takes account of **ZTU_MEM**, **ZTU_PC** and **ZANYMEM**), calculated across all private sector workplaces (**ASTATUS**<3) tells us that, on average, 10.9 per cent of employees in private sector workplaces are union members. This is based on an (unweighted) sample size of 1479. Again, we wish to know how reliable this estimate is (in other words, what it enables us to say about the population).

As noted above, the formula for the standard error of a proportion under SRSWR is as follows:

$$s.e.(\bar{x}) = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n^2}} = \frac{s.d.(x)}{\sqrt{n}}$$

Where *s.d.(x)* is the standard deviation of *x*. Again, we ignore the finite population correction, for simplicity.

The standard deviation of our union density variable in the private sector is 23.0. The SRSWR standard error of our sample mean of is therefore 0.60. So, under SRSWR, we could be 95 per cent confident that the mean union density *in the whole population* of private sector workplaces lies between 9.7 per cent and 12.1 per cent (or between 10 per cent and 12 per cent, after rounding).

However, Table 8A of the WERS98 Technical Report shows that **NDENSITY** has a design factor of 1.37. The true standard error of our density estimate under the WERS98 sample design is therefore 1.37*0.6=0.8. Accordingly, we can actually only be 95 per cent confident that the true population value lies between 9.3 per cent and 12.5 per cent (or between 9 per cent and 13 per cent after rounding). This is the true measure of the reliability (precision) of our estimate of 11 per cent.

4.6.2 Tabular analysis

By tabular analysis we mean analysis that aims to either:

a) compare estimates for different types of workplace, to see if the incidence varies across different parts of the population
b) examine the relationship between two categorical variables in order to test their independence.

First, consider (a):

Running a weighted table of **IPOLICY** by **ASIC** tells us that 61 per cent of Wholesale and Retail establishments have a formal written equal opportunities policy, compared with 71 per cent of those in the Hotel and Restaurant sector. The percentages are based on (unweighted) sample sizes of 320 and 126 respectively. We

wish to know whether our estimates are reliable enough to say that a difference also exists between the two groups in the population as a whole.

The test is based on the principle that, just as estimates have a confidence interval, so does the number representing the difference between the estimates. In our example, we are questioning whether we can be confident that the difference is not zero in the population as a whole.

The formula for the standard error of a difference between two proportions is as follows:

$$s.e.(p-q)=\sqrt{\frac{p(1-p)}{n_1}+\frac{q(1-q)}{n_2}}$$

The SRSWR standard error of our difference of 10 per cent is 4.9. So, under SRSWR, we could be 95 per cent confident that, *in the population as a whole*, the incidence of equal opportunities policies is higher within Hotels and Restaurants than it is within Wholesale and Retail. But only just. We would perhaps be more comfortable saying that we can be 90 per cent confident.

However, as seen in the previous section, **IPOLICY** has a design factor of 1.9. Multiplying the standard error by 1.9 gives a true standard error of 9.3. With this standard error, the test fails at both the 95 per cent and 90 per cent levels of confidence.

Next, consider (b):

The common test of independence between two categorical variables uses the Pearson chi-squared measure :

$$X_P^2 = n\sum_{r=1}^{R}\sum_{c=1}^{C}\frac{(\hat{p}_{rc}-\hat{p}_{0rc})^2}{\hat{p}_{0rc}}$$

where: *n* is the total number of observations, $\hat{p}_{rc}$ is the estimated proportion for the cell in the *r*th row and *c*th column of the table, and $\hat{p}_{orc}$ is the estimated proportion under the null hypothesis of independence.

Under SRSWR, this statistic is distributed asymptotically as chi-squared with *(R-1).(C-1)* degrees of freedom. However, under complex sample designs, the statistic is no longer distributed in this way (Rao and Thomas, 1989). The value of the standard test statistic will not, therefore, give a valid measure of the independence of the two variables. Its values will generally be too large, leading you to reject the null hypothesis of independence on occasions when this conclusion is not justified.

The preferred means of correcting the statistic is considered to be the second-order Rao-Scott correction (Sribney, 1998). This correction turns the Pearson chi-squared statistic into an F statistic with non-integer degrees of freedom. The correction is computationally very complex but, fortunately, it is available within STATA, where it

appears as the default test statistic on the `svytab` command. Here, the test gives you an adjusted significance level that can be used in the same way as the significance level that would otherwise be produced by the standard chi-squared test.

Unfortunately, there does not appear to be a similar correction available within SPSS.

4.6.3 Regression analysis

We saw in Section 4.6.1 that complex sample designs, such as that used in WERS98, lead to larger standard errors and wider confidence intervals in univariate analysis (frequencies) than are implied by SRSWR procedures. This is also true in regression analysis (Pfefferman, 1996; Skinner, 1989a, 1989b). As a result, users conducting regression analysis of data from WERS98 must also take account of the sample design in some way. This can be done either through 'aggregated' or 'disaggregated' methods. Aggregated methods involve constructing the model in the normal way, but using special techniques to adjust the standard errors. Disaggregated methods make the necessary adjustments by incorporating terms in the model that account for the sample design.

Aggregated methods

In these methods, the regressions are run on weighted data in order to obtain regression coefficients that are not biased by the unrepresentative nature of the sample. Special techniques are then employed to account for the sample design in the estimation of standard errors and confidence intervals. It should be noted that standard inference procedures, such as the Likelihood-Ratio test and residuals analysis are rendered invalid under these methods (Pfefferman, 1996: 252).

Skinner (1989b) suggests three different aggregated methods. They are listed here in order of the ease with which they may be applied by users with access to the standard versions of STATA and SPSS.

(i) Use a variance estimation technique that is robust to complex sample designs

Skinner (1989b: 78-79) derives a linearized variance estimator that accounts for complex sample designs. If an estimator of this type is employed by the regression procedure, the non-SRSWR nature of the sample will be taken into account in the calculation of the standard errors.

The variance estimator derived by Skinner (called a 'robust variance estimator' in the STATA manuals) is automatically called by STATA's `svy` estimators (e.g. `svyreg`, `svylogit`). The same variance estimator is also called when pweights are specified on non-`svy` estimation commands. But the `svy` commands make additional adjustments to the standard errors to account for stratification and clustering, and also make finite population corrections, as long as these items are specified on `svyset`, along with the weight, prior to the estimation – see Section 4.5.2. Further differences between the `svy` and non-`svy` commands are listed on pages 331-2 of the STATA User Guide.

For those with access to STATA, we would recommend use of the svy family of commands as the most straightforward means of accounting for the WERS98 sample design when conducting regression analysis. Unfortunately, SPSS does not include a 'linearized' variance estimator that is robust to complex sample designs. An alternative for SPSS users would be to adjust the SRSWR standard errors using an estimated design factor (*deft*), as described below.

(ii) Adjust the SRSWR-based standard errors using an estimated *deft*.

In this second method, the analyst should first run a weighted regression to obtain unbiased coefficients. The analyst should then run an unweighted regression to obtain SRSWR standard errors. The SRSWR standard error of each coefficient should then be multiplied by the *deft* of the mean of the dependent variable. Skinner states that this will usually give a conservative (sometimes over-conservative) estimate of the true standard error under the complex design (Skinner, 1989b: 77).

However, users should note Skinner's recommendation that the unweighted regression used to produce the uncorrected standard errors should employ a variance estimator which produces a heteroscedasticity-robust SRSWR standard error. This is because heteroscedasticity can bias standard errors even more than complex sample designs (Skinner, 1989b: 77). Such an estimator is variously referred to as the Huber, White, 'sandwich' or 'SRS linearized' estimator. It can be used to produce heteroscedasticity-robust standard errors without the user having to specify the precise nature of the heteroscedasticity, as you would under Weighted Least Squares.[10]

This approach of adjusting the SRSWR standard errors using an estimated *deft* may prove attractive to SPSS users, who are unable to follow option (i). However, to our knowledge, SPSS does not include a variance estimator that produces a heteroscedasticity-robust SRSWR standard error. SPSS users should therefore also take care to test and correct for heteroscedasticity where possible.

Given that STATA incorporates a variance estimator that is robust to complex sample designs, as outlined in option (i), this second approach is unlikely to prove attractive to STATA users.

(iii) Use 'replication' methods

Replication methods involve selecting sub-samples from the full sample, computing the desired statistic within each sub-sample, and then using the variability among the sub-sample estimates to compute the standard error of the full sample estimate. The sub-samples are called replicates, hence the term replication. Skinner (1989a: 51-5) mentions a number of replication methods including balanced repeated replication, the jackknife approach and bootstrapping.

Replication methods are not currently supported by the either SPSS or STATA. However, we will be investigating a piece of software called WESVAR that can reportedly be used in conjunction with SPSS to compute replicate variance

[10] The linearized estimator for complex designs, discussed under option (i), is an extension of this SRSWR estimator.

estimates.[11] STATA reports that it may incorporate replication methods as alternatives to the 'linearized' variance estimator in future versions of its `svy` commands.

We have not used any of these methods ourselves and so we are currently unable to comment any further on their use with WERS98 data. Skinner (1989a: 54) notes that none of the replication methods performs uniformly best across all statistics, designs and populations, and so we will be consulting reviews of the methods, such as those by Rust (1985) and Wolter (1985, Chapter 8), in order to assess their relative performance with WERS98-type data. Our conclusions will appear in a subsequent version of this Guide. However, for the time-being, we note Brick and Morganstein's comment that jackknife methods are likely to be the most appropriate for establishment surveys in which unequal sampling fractions are employed within different strata (Brick and Morganstein, n.d.).

Disaggregated method

This method involves estimating an unweighted regression in which the sample design is fully accounted for by including variables that describe the sample design as covariates in the model.

The advantage of this method is that standard SRSWR-based inference methods can still be used (Pfefferman, 1996: 255). However, there are potential drawbacks. The first is that the information is not yet available to be able to specify covariates that fully account for the sample design (although, as shown below, this may not matter). The second is that the user may not feel it appropriate to include a large number of additional variables in the model if they are not of direct scientific interest (although this is not a problem if one is merely aiming for the greatest level of explanatory power from the model).

Disaggregated analysis of data from the 1990 Workplace Industrial Relations Survey was the subject of an unpublished paper by Chris Skinner from the University of Southampton (Skinner, 1997). Here, we attempt to extend his recommendations to cover the workplace data from WERS98.[12] Our general conclusion would be that the nature of the sample design makes disaggregated analysis of the WERS98 workplace data a formidable task. However, we explain the method here so that users are aware of the possibilities and the apparent pitfalls.

The principal sample design feature that needs to be taken account of in the workplace data from WERS98 is the use of variable sampling fractions within different strata. This can be accounted for by including dummy variables that identify workplaces arising from the same stratum on the sampling frame. A variable that groups workplaces arising from the same stratum (**IDBRSTR1**) is available on the restricted data file **Sample98.sav**. The variable has 72 categories, which can be converted into dummies for inclusion in the model. Adding 71 of these dummies to the list of covariates will remove the major source of selection bias in the model coefficients (i.e. the use of unequal sampling fractions). However, when incorporating the

[11] Further details of the software are available at: http://www.westat.com/wesvar/index.htm

[12] The employee data contains further complexity because of the clustering of employees within workplaces and so is not considered here. It may form part of a subsequent version of this Guide.

dummies, one must explore possible interactions with other variables in the model in case there are different regression slopes in different strata (Skinner, 1989: 215, 1997).

There remains the possibility that selection bias may also have resulted from the differential probability of sampling for establishments corresponding to different numbers of census units (as described in Section 7.1.1 of the Technical Report). In these cases, the dummy variables describing the stratification of the sample will not account for the non-standard probability of selection.

Unfortunately, information is not yet available to permit users to adjust the model to take account of these non-standard cases. One must therefore consider whether the fact of an establishment having a non-standard probability of selection is likely to be related to the values of the dependent variable, after controlling for all other factors in the model. If the two are unrelated, then the non-standard probability of selection of these cases introduces no bias into the model coefficients (as it is unrelated to the error term) and can be ignored.

One can attempt to check this by comparing weighted and unweighted estimates produced by the model (including the stratum variables in both the weighted and unweighted case). If the non-standard probabilities are not biasing the coefficients, all that might be observed is an inflation of standard errors and corresponding 'random variation' in the coefficients (Skinner, 1997). The hypothesis that the difference between the weighted and unweighted estimates is merely due to sampling variation can be formally tested, using methods outlined by DuMouchel and Duncan (1983) or Pfefferman (1993). However, if some systematic difference is observed, there are four possibilities:

(i) The effects of the stratification dummies have not been accurately specified in the model.
(ii) The non-standard probabilities of selection are generating some selection bias to the coefficients of the unweighted model.
(iii) The trimming of extreme weights (Airey *et al.*, 1999: 90) means that, although (i) and (ii) are not true, the weighted and unweighted estimates are still systematically different, since the weights do not accurately reflect the true probability of selection.
(iv) There remains some unexplainable misspecification.

We hope to be able to make available variables that: firstly, identify those cases with non-standard probabilities of selection; and secondly, provide an untrimmed weight. Until those variables are available, one could perhaps only confidently pursue this disaggregated approach if one is willing to assume the following:

(i) In compiling the disaggregated model, the user has included terms that fully specify the effects of the sample stratification (possible involving interactions with other variables in the model)
(ii) The non-standard probabilities of selection do not introduce any selection bias
(iii) The trimming of extreme weights does not affect comparisons between unweighted and weighted estimates from the disaggregated model.

Even so, the fact of needing to include at least 71 dummies to account for the stratification would seem to be a significant obstacle to those considering this approach.

5. The production of high quality tables in SPSS

5.1 Introduction

This section of the Introductory Guide aims to provide a quick guide to the SPSS Tables module, focusing on those elements of SPSS Tables that you can use to produce high quality tabular analysis of WERS98 data.

SPSS Tables is an 'add-on module' to the SPSS Base system. It provides greater control over the content and appearance of tables when compared with the standard SPSS `crosstabs` and `mult response` commands or their equivalents in STATA. Specific advantages over the standard SPSS commands include:

- more accurate calculation of proportions from weighted data (see Section 4.5.1)
- considerable flexibility in the presentation of statistics
- the ability to include weighted and unweighted figures on the same table – an extremely helpful facility since unweighted bases help you to gauge the precision of your estimates.

These various features made SPSS Tables an invaluable tool in the primary analysis of WERS98. This section is intended to pass on some of the valuable techniques used during that analysis.

You can check whether the SPSS Tables module is already installed on your system by starting SPSS and pulling down the 'Analyze' menu on the SPSS toolbar. If SPSS Tables is installed you will see an option labelled 'Custom Tables' on this menu, under that labelled 'Descriptive Statistics'. If SPSS Tables is not installed, you should contact your system administrator. All of the following discussion concerns SPSS Tables version 8.0.

5.2 Preparation

A few preparatory tasks need to be carried out before you first use the SPSS Tables module.

- First, you need to decide on your preferred style of table. This choice governs the appearance of your tables (e.g. line style, cell shading and the like) not the content, which will be determined later. Pull down the 'Edit' menu from the SPSS toolbar and select the 'Pivot Tables' tab from the 'Options' menu. A list of 'TableLooks' should be displayed, beginning with <System default> and continuing through ACAD2VGA.TLO and ACADEM2.TLO to VERTIME.TLO.[13] Scroll through the list and choose the style of table that you prefer.
- Now select the 'Output Labels' tab from the top of the 'Options' window. Use the second pull-down menu under the heading 'Pivot Table Labelling' to determine

[13] If only <System default> is displayed, highlight this 'TableLook' and click on the button labelled 'Browse…'. A new dialog will be displayed, which you should cancel. The full list of 'TableLooks' should now appear in the initial window.

whether the tables you will produce should contain values, value labels or both. (A bug means that SPSS Tables seems always to display variable labels, irrespective of which setting is chosen in the first pull-down menu).

These things only need to be done once, not at the start of each session.

5.3 Basic table specification

The basic specification of a Tables command is outlined below using syntax and menus. In more complex specifications, covered in subsequent sections, the menu-based procedures are shown to be less flexible than the syntax-based route. However, both options are given in each case for completeness.

The output is best displayed in the Output Viewer rather than the Draft Viewer. In the examples referred to below, the TableLook is set to ACDEMIC.TLO, with value names and labels shown on the table.

Using syntax:

The syntax required to produce specify a simple table is reproduced in Example 1 below. This syntax first reads in the data (in this and all other examples: the final version of the WERS98 Management data in SPSS format, **Mq98fin.por**). It then weights the data by the workplace-level weighting variable and then produces a table of **EANYEMP** (dichotomous variable indicating whether any employees belong to trade unions) by **NEMPSIZE** (categorical variable indicating size of workforce). The output is headed Example 1 in Appendix E.

The syntax may look rather daunting when compared with the `crosstabs` command, but once you have found a specification that you are happy with, it can be quickly and easily extended to produce further tables. Each element of the syntax is described below.

Example 1:

```
import file="D:\WERS98\Mq98_fin.por" .
weight by est_wt .
tables
   /format blank missing('.')
   /ftotal base1 'Base' total1 ‘All w/places’
   /autolabel=on
   /missing=include
   /base=qualified
   /table (eanyemp + base1) by (nempsize + total1)
   /statistics cpct(eanyemp(f3) '':nempsize)
    count(base1 ‘Weighted') u count(base1 'Unweighted') .
```

The `format` command controls the appearance of certain types of cell. Here, the `blank` statement specifies that empty cells which would otherwise contain counts or percentages should be left blank (rather than containing a 0, for which `zero` should

be used). If `blank` is specified, the appearance of 0% in a cell would therefore mean "a non-zero value less than 0.5%", rather than absolute zero. The `missing ('.')` statement does the same for empty cells which should otherwise contain summary data (such as means), here specifying that they should contain a 'period'. The alternative is `missing 'chars'`, where *`chars`* might be the word 'Missing' or a symbol such as '$'.

`ftotal` sets up two elements, `base1` and `total1`, which are "following totals" (i.e. totals that will follow a chosen variable in either a row or column of the table). `base1` will be used as a base element and tacked onto the bottom of the row variable **EANYEMP**, where it will appear with the label 'Base'. `total1` will be used as a summary column and tacked onto the end of the column variable **NEMPSIZE**, where it will appear with the label 'All w/places'.

`autolabel=on` automatically prints a default table title (consisting of the contents of the `table` subcommand). The alternative is `autolabel=off`.

`missing=include` specifies that user-missing values should be included in the table, although there are no user-missing values on either variable in our table. The alternative is `missing=exclude`.

`base=qualified` typically accompanies `missing=include` and specifies that user-missing values should be treated like other values in the calculation of percentages or summary statistics. `base=all` includes user and system-missing values; `base=answering` excludes all missing values. If one does not wish to include missing values in the table, one should simple delete the `missing` and `base` rows from the table specification, since `missing=exclude` and `base=answering` are the default settings.

The `table` subcommand gives the specification of the table itself. Here the `base1` element is tacked onto the bottom of **EANYEMP** using the "+" sign and then the combined axis is tabulated by **NEMPSIZE** (which itself has `total1` tacked on to it).

The `statistics` subcommand controls the contents of each of the cells of data in the table and is the most complex part of SPSS Tables. Taking it piece by piece:

`cpct(eanyemp`	Specifies that column percents should appear in those rows relating to the variable **EANYEMP** (and so not in those rows relating to the `base1` element)
`(F3) ''`	Specifies that 3 digits should be allowed for these column percents. F3.1 would also permit one decimal place to be printed; here we are printing only integers. The alternative `pct4` format would add a "%" symbol in the additional column as the end of each value. The " after the closing bracket stops the label "CPCT" appearing after the value label on each row.
`:nempsize)`	Specifies that the column percentage should be calculated through dividing the cell count by the total number of cases within each value of NEMPSIZE

(omitting it would cause the cell count to be divided by the total number of cases in the table, i.e. across all values of NEMPSIZE).

`count(Base1 'Weighted')` Specifies that counts should appear on the `base1` element. (If the data is weighted, these will be weighted counts). It also specifies that this count element should be labelled "Weighted".

`u count(Base1 'Unweighted')` Specifies that *unweighted* counts should also appear on the `base1` element, and that this row should be labelled "Unweighted".

Additional tables can be produced to the same general specification by simply replicating the last two rows of the specification. The first five subcommands in the `tables` syntax in Example 1 (`format, ftotal, autolabel, missing` and `base`) are all *global* subcommands and will apply to all tables subsequently specified on that single `tables` command. The last two subcommands (`table` and `statistics`) are *local* subcommands and can be repeated, as follows:

Example 2:

```
tables
  /format blank missing('.')
  /ftotal base1 'Base' total1 'All w/places'
  /autolabel=on
  /missing=include
  /base=qualified
  /table (eanyemp + base1) by (nempsize + total1)
  /statistics cpct(eanyemp(f3) '':nempsize)
   count(base1 'Weighted') u count(base1 'Unweighted')
  /table (aphras01 + base1) by (astatus + total1)
  /statistics cpct(aphras01(f3) '':astatus)
   count(base1 'Weighted') u count(base1 'Unweighted').
```

The additional output produced by the second command is presented in Appendix E.

Using the menu system:

The table pictured in Example 1 of Appendix E can equally be produced using the SPSS menu system, as follows:

1. Open the Management data set and weight the data by **EST_WT** (see Section 4.4 above).
2. From the 'Analyze' pull-down menu, select 'Custom Tables'. Select 'General Tables' from the new menu.
3. Highlight the variable **EANYEMP** in the variable list and use the arrow button to transfer the variable into the list titled 'Rows:'.
4. Click on the button labelled 'Edit Statistics' to determine the cell statistics for these rows. Select 'Col %' from the list and click on the button labelled 'Add' to

move it into the 'Cell Statistics' list. Remove any other elements, such as 'Count'. Then highlight 'Col %' and adjust the Format to 'ddd.dd' using the pull-down menu. Adjust the 'Width' to 3 and the 'Decimals' to 0. Delete the Label 'Col %'. Then click on the button labelled 'Change', followed by 'Continue'.

5. Insert a base element for the row variable by clicking on the button labelled 'Insert total'. A total named '**eanyempTotal**' will be added to the 'Rows' list. Highlight the name and click on 'Edit Statistics'. When the new window appears, check 'Custom total statistics' at the top. Then add 'Count' and 'Unweighted N' to the 'Cell Statistics' list. Click on 'Continue'.
6. Repeat a similar procedure to transfer the variable **NEMPSIZE** into the list titled 'Columns:' and insert a following total '**nempsizeTotal**'. Note, however, that you will not be able to edit the Statistics for these elements as you have already determined the statistics to be printed in the table.
7. Use the button labelled 'Formats' to display the FORMAT options – described above under *Using syntax*. Titles can be set using the button labelled 'Titles', although there is no facility for setting AUTOLABEL when producing tables using the menus.
8. Click on 'OK' to run the table.

The table produced by this menu-based procedure is exactly the same as that produced by the syntax outlined in the previous section, except for the absence of customized labels on certain items such as the following totals. However, we have found the syntax-based method to be preferable, particularly because of the ease with which additional tables can be added to the specification using Copy and Paste.

Users should also note that user-missing values are automatically excluded from tables produced using the menu system. There does not appear to be any facility for including them, as there is when using syntax.

5.4 More complex specifications

A variety of more complex tables can be specified using either syntax or menus. These are outlined below.

5.4.1 Summarising continuous variables

Using syntax:

The `tables` command needs to be amended for producing tables of means, medians and the like. First, the `missing` and `base` subcommands are removed and a new global subcommand is inserted: `observation`. This identifies the continuous variable whose values we wish to summarise in the table. Here we wish to look at mean percentage of days lost to employee absence (**ZABSENCE**) within each category of **ASTATUS**. The `cpct` element of the `statistics` subcommand is replaced by `mean`, with `f3.1` indicating that 4 columns will be sufficient to display the results, one following the decimal place. The `count` and `u count` elements are replaced with `validn` and `u validn` respectively, which count the number of non-missing values of an observation variable.

Example 3:

```
tables
 /format blank missing('.')
 /ftotal base1 'Base' total1 'All w/places'
 /autolabel=on
 /observation=zabsence
 /table (zabsence + base1) by (astatus + total1)
 /statistics mean(zabsence(f3.1))
  validn(base1 'weighted') u validn(base1 'unweighted') .
```

Two single quotes `''` could be inserted after `(F3.1)` in order to remove the 'Mean' label shown in the output in Appendix E.

Using the menu system:

1. Follow steps 1 and 2 outlined in Section 5.3
2. Highlight the variable **ZABSENCE** in the variable list and use the arrow button to transfer the variable into the list titled 'Rows:'.
3. To the right of the 'Rows:' list, under the heading 'Selected Variable', check the option labelled 'Is summarized'.
4. Click on the button labelled 'Edit Statistics' to determine the cell statistics. Select 'Mean' from the list and click on the button labelled 'Add' to move it into the 'Cell Statistics' list, if it is not already there. Remove any other elements. Then highlight 'Mean' and adjust the Format to 'ddd.dd' using the pull-down menu. Adjust the 'Width' to 3 and the 'Decimals' to 1. Then click on the button labelled 'Change', followed by 'Continue'.
5. Transfer the variable **NEMPSIZE** into the list titled 'Columns:' and insert a following total '**nempsizeTotal**'. You will not be able to edit the Statistics for these elements as you have already determined the statistics to be printed in the table.
6. Use the button labelled 'Formats' to display the FORMAT options – described above under *Using syntax*. Titles can be set using the button labelled 'Titles', although there is no facility for setting AUTOLABEL when producing tables using the menus.
7. Click on 'OK' to run the table.

Note that we have been unable to find a means of adding a Base element to these types of tables via the menu system.

5.4.2 Aggregating continuous variables

In some cases, analysts may wish to produce an aggregate measure of a continuous variable across all workplaces in a particular sector. A common use of this technique in past WIRS source books has been the analysis of union density. Commonly, the source books have calculated the overall percentage of employees that are union members *across* a set of workplaces (as per Table 10.11 in Millward *et al*., 1999), in addition to calculating the average density *within* workplaces (as per Table 10.10, *ibid.*). The calculation of the latter (mean workplace density) is possible using the

procedure outlined in Section 5.4.1 above. The calculation of an aggregate measure requires only a minor amendment to that method.

This amendment merely involves calculating a new weight variable, equal to the existing weight multiplied by the number of employees in the workplace, and then running the syntax or menu procedure from Section 5.4.1 under the new weighting system. To illustrate, we return to **ZABSENCE** since a derived variable for union density is not immediately available on the data file. We present only the syntax, omitting the menu-based alternative, since the essential change from Section 5.4.1 is in the compilation and application of a new weight, and not in the form of the tabulation procedure.

Example 4:

```
compute eeweight=zallemps*est_wt .
weight by eeweight .
tables
 /format blank missing('.')
 /ftotal base1 'Base' total1 'All w/places'
 /autolabel=on
 /observation=zabsence
 /table (zabsence + base1) by (nempsize + total1)
 /statistics mean(zabsence(f3.1))
  validn(base1 'weighted') u validn(base1 'unweighted') .
```

This procedure gives slightly different results to those given by Example 3. Whereas Example 3 showed us that the mean percentage of days lost to employee sickness in public sector workplaces was 5.4, Example 4 shows us that, overall, 5.0 per cent of public sector work days were lost to employee absence.

Analysis of union density generally gives starker differences because union membership is more unevenly distributed between small and large workplaces than is absence.

5.4.3 Multiple-response items

WERS98 includes numerous multiple-response questions (i.e. questions where interviewers may record more than one response from the interviewee). The question from which the variables **BYOURJ01-BYOURJ10** derive is an example. Here, up to 10 responses were recorded by the interviewer. (In fact, a maximum of 9 responses were received and so **BYOURJ10** has been dropped from the data file.)

If one wishes to assess the incidence of the various job responsibilities recorded on **BYOURJ01-BYOURJ09**, there are three options. Either: produce nine separate tables and add up the incidence of each responsibility across the nine tables; produce a new dichotomous variable which is true if a particular responsibility has been mentioned in any one of the nine variables; or run a composite table which will automatically compile the information on a single table. The following procedure shows how the last option can be achieved within SPSS Tables.

Using syntax:

The basic syntax command needs to be added to by specifying a new 'temporary' variable that groups the multiple-response items together. (The variable is temporary in the sense that it is not available to SPSS procedures other than the `tables` command which defines it.) The temporary variable is defined by using the `mrgroup` subcommand. This variable is then tabulated in the normal way using the standard syntax outlined in Example 1.

There are three elements to the `mrgroup` subcommand: a user-defined name for the temporary variable (here 'byourj'), a user-defined label for the temporary variable (here 'Work responsibilities of respondent and their subordinates') and a list of the variables containing the multiple-response items (here **BYOURJ01** to **BYOURJ09**).

Example 5:

```
tables
   /format blank missing('.')
   /ftotal base1 'Base' total1 'All w/places'
   /autolabel=on
   /missing=include
   /base=qualified
   /mrgroup=byourj 'Work responsibilities of respondent
    and their subordinates' byourj01 to byourj09
   /table (byourj + base1) by (nempsize + total1)
   /statistics cpct(byourj(f3) '':nempsize)
    count(base1 'Weighted') u count(base1 'Unweighted') .
```

The output from Example 5 (contained in Appendix E) shows that 79 per cent of respondents in small workplaces (10 to 24 employees) reported that pay and conditions formed part of their own work responsibilities, or the work responsibilities of their subordinates.

Using the menu system:

1. Follow steps 1 and 2 outlined in Section 5.3
2. When the 'General Tables' dialog box appears, click on the button labelled 'Multiple Reponse Sets…' in the bottom left hand corner of the window.
3. From the list of variables headed 'Set Definition', select the 9 variables **BYOURJ01** to **BYOURJ09** and use the arrow button to transfer them into the list headed 'Variables in Set:'.
4. Under the heading 'Variables Are Coded As', check the button labelled 'Categories' (as opposed to 'Dichotomies').
5. Give the multiple response variable a Name (e.g. BYOURJ) and a Label (e.g. 'Work responsibilities of respondent and their subordinates').
6. Ensure that the 'Denominator for Multiple-Response Percentages' is selected as 'Number of cases' (as opposed to 'Number of responses').
7. Click on the 'Add' button, followed by the 'Save' button. The temporary multiple-response variable, labelled **$byourj** should now appear in the list at the bottom left-hand corner of the General Tables window.

8. Highlight the **$byourj** variable and use the arrow button to transfer the variable into the list titled 'Rows:'.
9. Click on the button labelled 'Edit Statistics' to determine the cell statistics for these rows. Select 'Col Response %' from the list and click on the button labelled 'Add' to move it into the 'Cell Statistics' list. Remove any other elements, such as 'Respondents'. Then highlight 'Col Response %' and adjust the Format to 'ddd.dd' using the pull-down menu. Adjust the 'Width' to 3 and the 'Decimals' to 0. Delete the Label 'Col Response %'. Then click on the button labelled 'Change', followed by 'Continue'.
10. Insert a base element for the row variable by clicking on the button labelled 'Insert total'. A total named '**$byourjTotal**' will be added to the 'Rows' list. Highlight the name and click on 'Edit Statistics'. When the new window appears, check 'Custom total statistics' at the top. Then add 'Respondents' to the 'Cell Statistics' list. Click on 'Continue'. This will provide the weighted number of cases in the base element. Note that it does not appear possible to add the unweighted number, as it is in the basic table specification outlined in Section 5.3.
11. Transfer the variable **NEMPSIZE** into the list titled 'Columns:' and insert a following total '**nempsizeTotal**'. Note, however, that you will not be able to edit the Statistics for these elements as you have already determined the statistics to be printed in the table.
12. Use the button labelled 'Formats' to display the FORMAT options – described above under *Using syntax*. Titles can be set using the button labelled 'Titles', although there is no facility for setting AUTOLABEL when producing tables using the menus.
13. Click on 'OK' to run the table.

5.5 Final notes

The examples given throughout Section 5 above should cover most types of table that you will need to produce in your analysis. In most cases, therefore, users will be able to follow the syntax or menu instructions given above and simply change the variable names as appropriate.

If situations arise in which users wish to produce a particular type of table not shown above, they are referred to the Syntax Guide in the back of the User Manual for SPSS Tables 8.0 or the on-line Help in SPSS, both of which give further assistance.

6. Combining data from separate files for linked analysis

There are a number of different reasons why users may wish to combine data from separate files in WERS98. For instance, users may wish to:

1. Combine data from the Management data file with that from the Worker Representative data file in order to compare responses from managers and worker representatives within the same workplace (e.g. on issues such as the incidence of industrial action, using **GACTIO01-04** and **WHINDA01-04**)
2. Add data from the Management or Worker Rep data files onto the Employee data file in order to be able to distinguish employees according to the characteristics of their workplace (e.g. size or industry).
3. Produce summary information about the workforce in an establishment from the records in the Employee data file (e.g. average levels of job satisfaction) and then use this in combination with workplace-level data from the Management or Worker Representative data files
4. Combine data from the Management, Worker Representative, Employee or Panel data files with verbatim responses contained in the Excel spreadsheets.

Each of these four tasks can be accomplished in SPSS or STATA with the minimum of effort, once one is familiar with the necessary commands. This section aims to show how this may be done.

We do not, however, seek to say a great deal about how the resulting data files may be analysed. Options 2 and 3 above generate linked employer-employee data that will be relatively new to most users. Analysis of this data therefore provides new opportunities, but also some new problems, particularly for those wishing to use econometric methods. We address one of these problems in Section 6.3.3; namely, the issue of generalizability when producing summary data under Option 3. For further guidance on the econometric analysis of linked employer-employee data, readers are referred to Haltiwanger *et al.* (1999).

6.1 Combining data from the Management and Worker Representative data files

The Management and Worker Representative data files are both workplace-level data files. Each and every workplace that participated in the WERS98 Cross-Section Survey has a single record in the Management data file. A selection of these workplaces (namely, those in which eligible Worker Representatives were present and participated in the Survey) also have a single record in the Worker Representative data file. The process of combining data from the Management and Worker Representative data files therefore involves a *'one-to-one' match*, so-called because one record from the first data file is matched with one (and only one) other from the second data file. The alternative – a *'one-to-many' match* – is discussed in Section 6.2.

This matching process (referred to as 'merging' in STATA) is made possible by the fact that each workplace in WERS98 has its own unique identifier (**SERNO**), which is present on both of the files. Combining data from the two files therefore simply involves combining cases with matching values on the **SERNO** variable.

Since the match is 'one-to-one', that match can take place in either direction. In other words, you can match the Worker Representative data onto the end of the Management data file or, alternatively, you can match the Management data onto the end of each Worker Representative record.

Option A:

If one wishes to obtain a data file containing all of the Management records, with Worker Representative data present wherever they were interviewed, then one needs to match the Worker Representative data onto the Management data file. The resultant data file will look something like this:

SERNO		
1	Manager's data	
2	Manager's data	Worker Representative's data
3	Manager's data	
4	Manager's data	Worker Representative's data

etc.

In SPSS terminology, the Management data file is here referred to as the 'working data file' whilst the Worker Representative file is referred to as the 'lookup data file' (since one is initially working with the Management data and then 'looks up' relevant cases from the Worker Representative file). In STATA, they are referred to as the 'master data file' and the 'using' data file respectively (since you perform a merge onto the Management data 'using' the Worker Representative data).

Option B:

If, on the other hand, one wishes to obtain a data file containing only those workplaces in which Worker Representatives were interviewed, with the relevant Management data added on, then one needs to match the Management data onto the Worker Representative data file. The resultant data file will then look something like this:

SERNO		
1	Worker Representative's data	Manager's data
2	Worker Representative's data	Manager's data
3	Worker Representative's data	Manager's data
4	Worker Representative's data	Manager's data

Etc.

In this case, the Worker Representative data file is referred to as the 'working data file' (or 'master data file' in STATA) and the Management file is the 'lookup data file' (or 'using data file').

Under both options, the resultant data file contains workplace level data. Accordingly, the combined data is weighted by **EST_WT**: the standard workplace level weight.

6.1.1 Combining the data in SPSS

The matching of the two data files in SPSS is achieved by using the `match files` command. The necessary syntax and menu-based procedures are set out below.

Before proceeding, however, users should note that `match files` will only work with files saved in ***.sav** format. The SPSS WERS98 data is generally supplied in ***.por** format. These files therefore need to be converted to ***.sav** format before the `match files` can be used through either the syntax or menu-based route.

Users should also note that the `match files` command requires that both data files are sorted in ascending order of the key variable (**SERNO** in this case). The Management and Worker Representative data files are sorted in this way when supplied by the Data Archive. However, if users wish to use the command with data files that they have themselves derived from the source files (or if they have re-sorted and saved the original files), they must ensure that the two files are sorted by **SERNO** before matching.

Using syntax:

The required syntax is as follows:

```
match files file="d:\wers98\mq98fin.sav"
   /table="d:\wers98\wrq98.sav"
   /by serno .
```

If **Mq98fin.sav** is already open in the SPSS Data Editor, the phrase `"d:\wers98\mq98fin.sav"` can be replaced with an asterisk, as follows:

```
match files file=*
   /table="d:\wers98\wrq98.sav"
   /by serno .
```

In both of these examples, all of the variables in the Worker Representative data file will be matched onto the end of the appropriate record in the Management data file. Users are referred to the on-line SPSS User Manual for details of the additional functionality that is available from the `match files` command, such as the ability to keep and drop sets of variables during the matching process.

Using the menu system:

1. Open the Management data file **Mq98fin.sav**.
2. From the ‘Data’ menu, select the ‘Merge Files’ option and then the subsequent option to ‘Add Variables…’.
3. Select **Wrq98.sav** as the read file.
4. In the ‘Add Variables’ window, check the square box labelled ‘Match cases on key variables in sorted files’. Then check the circle underneath it labelled ‘External file is keyed table’.
5. Select **SERNO** from the list headed ‘Excluded variables’ and use the arrow to the left hand side of the list headed ‘Key variables’ to transfer **SERNO** into this list.

6. Click on 'OK'. You will be warned that the match will not work if the files are not sorted in ascending order of the key variable (**SERNO**). As long as you are sure that the files are sorted, you can click on 'OK'. The data will then be combined in a new working data file.

All of the variables in the Worker Representative data file will then be matched onto the end of the appropriate record in the Management data file. Users are referred to the on-line SPSS Help for details of the additional functionality that is available through the menu-based `match files` procedure, such as the ability to keep and drop sets of variables during the matching process.

6.1.2 Combining the data in STATA

The matching of the two data files in STATA is achieved by using the `merge` command, through a procedure that STATA calls a 'match merge'. The necessary syntax is set out below.

Before proceeding, however, users should note that `merge` will only work if both data files (the 'master' and 'using' data file) are sorted in ascending order of the key variable (**SERNO** in this case). Both **Mq98fin.dta** and **Wrq98.dta** are ordered by **SERNO** when supplied by the Data Archive, but this is not recorded in the piece of internal information that STATA refers to before matching the data files. So, users must open each data file in turn and run the command `sort serno` to sort the data by **SERNO**, then save the data file again. This ensures that STATA 'knows' that the data is sorted by **SERNO**, so that it will let you run the `merge` procedure. You can check whether STATA 'knows' how the data is sorted by entering the describe command. At the bottom of the output will appear

`Sorted by:` if STATA does not know how the data is sorted, or
`Sorted by: serno` if STATA knows that it is sorted by SERNO.

Once the data have been sorted, the two data files can be merged using the following syntax:

```
set memory 5000
use "d:\wers98\mq98fin.dta", clear
merge serno using "d:\wers98\wrq98.dta"
```

For further details about the `merge` command, including details of how to check that it has worked as intended, users are referred to the entry on `merge` in the STATA Reference manuals.

6.2 Adding workplace data to the Survey of Employees data file

The nature of the sampling procedure for the Survey of Employees was such that Employee questionnaires were distributed only in those workplaces where Management interviews had already taken place. Accordingly, each employee record

has an equivalent set of workplace-level data in **Mq98fin.*** (and **Wrq98.***, where Worker Representatives were interviewed).[14]

The process of adding data from the Management or Worker Representative data files to the Survey of Employees data file therefore involves a *'one-to-many' match*. It is so-called because one record from the Management or Worker Representative data files is matched onto many records (potentially, up to 25) in the Survey of Employees data file.

As set out in Section 6.1 above, this matching process (referred to as 'merging' in STATA) is made possible by the fact that each workplace in WERS98 has its own unique identifier (**SERNO**), which is present on each of the Cross-Section data files. Adding workplace data to the Survey of Employees data file therefore simply involves combining cases with matching values on the **SERNO** variable.

The resultant data file will then look something like this:

SERIAL	**SERNO**		
11	1	Employee 1 in Workplace 1	Data from Workplace 1
12	1	Employee 2 in Workplace 1	Data from Workplace 1
13	1	Employee 3 in Workplace 1	Data from Workplace 1
21	2	Employee 1 in Workplace 2	Data from Workplace 2
	Etc.		

Here, the Survey of Employees data file is referred to as the 'working data file' (or 'master data file' in STATA) and the workplace-level data file (Management or Worker Representative) is the 'lookup data file' (or 'using data file' in STATA).

The resultant data file contains employee-level data. Accordingly, the combined data is weighted by **EMPWT_NR**: the standard weight for the employee data.

6.2.1 Adding the workplace data in SPSS

The matching of the data files in SPSS is again achieved by using the `match files` command. The necessary syntax and menu-based procedures are set out below. The same conditions as set out in Section 6.1.1 regarding the format and sorting of the data files apply.

Using syntax:

The required syntax is as follows:

```
match files file="d:\wers98\seq98.sav"
   /table="d:\wers98\mq98fin.sav"
   /by serno .
```

[14] The exceptions are those employees from workplace 13068. This workplace was deleted from **Mq98fin.*** at the end of fieldwork without its employees being deleted from **Seq98.***. See the document of *Variable Notes to Accompany the Survey of Employees Dataset and Questionnaire*, available from the WERS98 Data Dissemination Service web-site (www.niesr.ac.uk/niesr/wers98/) .

If **Mq98fin.sav** is already open in the SPSS Data Editor, the phrase `"d:\wers98\mq98fin.sav"` can be replaced with an asterisk, as follows:

```
match files file=*
   /table="d:\wers98\mq98fin.sav"
   /by serno .
```

In both of these examples, all of the variables in the Management data file will be matched onto the end of the appropriate records in the Survey of Employees data file. **Mq98fin.sav** may of course be replaced with **Wrq98.sav** in either example.

In either case, this will create a very large data file (28,215 observations and over 1,000 variables, in the case where the Management data is added). It would therefore be wise to create a smaller version of the Management data file, containing only those variables of interest, before matching onto the Survey of Employees data file. Alternatively, users may make use of the `/keep` and `/drop` subcommands, which give control over the variables that are kept in the new data file. Users are referred to the on-line SPSS User Manual's entry on `match files` for further details of these subcommands.

Using the menu system:

1. Open the Survey of Employees data file **Seq98.sav**.
2. From the 'Data' menu, select the 'Merge Files' option and then the subsequent option to 'Add Variables…'.
3. Select **Mq98fin.sav** as the read file (or, alternatively, **Wrq98.sav**).
4. Follow steps 4 to 6 in Section 6.1.1 to complete the process.

All of the variables in the Management data file will then be matched onto the end of the appropriate records in the Survey of Employees data file.

As stated in the preceding section on syntax, this will create a very large data file (28,215 observations and over 1,000 variables, in the case where the Management data is added). It would therefore be wise to create a smaller version of the Management data file, containing only those variables of interest, before matching onto the Survey of Employees data file. Alternatively, users may exclude variables at Step 5 of the matching process by transferring variables from the list headed 'New Working Data File' into the list headed 'Excluded variables'.

Users are referred to the on-line SPSS Help for details of the additional functionality that is available through the menu-based `match files` procedure.

6.2.2 Adding the workplace data in STATA

As in Section 6.1.2, the matching of the data files in STATA is achieved by using the `merge` command, through a procedure that STATA calls a 'match merge'. The necessary syntax is set out below. The points made in Section 6.1.2 about the necessity of sorting the data files before using the `merge` command apply here also.

Note before proceeding, however, that simply adding all of the variables in the Management (or Worker Representative) data file onto the Survey of Employees data file will generate a very large data file (28,215 observations and over 1,000 variables, in the case where the Management data is added). STATA will need at least 35Mb of available memory in order to even create and hold this new file! It would therefore be wise to create a smaller version of the Management data file, containing only those variables of interest, before matching onto the Survey of Employees data file. We use a hypothetical data file of this type, which we have called **Mq98smal.dta**, in the following example.

Once the data files have been sorted, workplace data can be added to the Survey of Employees data files by using the following syntax:

```
set memory 5000
use "d:\wers98\seq98.dta", clear
merge serno using "d:\wers98\mq98smal.dta", nokeep
```

The use of the `nokeep` option on the `merge` command ensures that workplace records from **Mq98smal.dta** for which there are no corresponding employee records in **Seq98.dta** are ignored and not brought into the new file.

Wrq98.dta can naturally be used in place of **Mq98smal.dta** in order to add data from the Worker Representative data file.

For further details about the `merge` command, including details of how to check that it has worked as intended, users are referred to the entry on `merge` in the STATA Reference manuals.

6.3 Aggregating data from the Survey of Employees

In Section 6.2 above, a *one-to-many match* was used to add data about each workplace onto the records of each employee at that workplace who completed and returned an employee questionnaire. But suppose that, instead, we wish to match information about these employees onto the workplace-level data? This would constitute a *many-to-one match*, which is not possible within the matching procedures outlined in Sections 6.1 and 6.2 if we wish to end up with a workplace-level file. Simply stated, it is not possible to place 2, 3 or more employee records into the one space at the end of each workplace-level record without manipulating the data in some way in SPSS or STATA.

6.3.1 Aggregating data from the Survey of Employees in SPSS

The most straightforward means of aggregating the employee data in SPSS is by using the `aggregate` command to generate a workplace-level data file that contains summary information about the employees from that workplace who participated in the Survey of Employees (e.g. mean number of hours worked).[15]

[15] A second, more involved method, involves creating a workplace-level data file in which each of the (up to 25) employee records are placed horizontally, one after the other. The employee data in this

The `aggregate` command takes the Survey of Employees data file and creates a new data file in which there is one record for each workplace. In producing the file, the command can create a range of summary data items containing, for example, the mean value of a particular variable for employees in that workplace, the minimum or maximum value amongst those employees, or the sum of all values amongst those employees.

Suppose that we wished to create a workplace-level data file containing three summary data items from the Survey of Employees: first, the mean number of hours worked by the participating employees in each workplace; second, the number of employees giving a valid (non-missing) response to the question on hours; and third, the total number of participating employees in each workplace. This workplace-level data file could then be matched onto the Management or Worker Representative data files using the method outlined in Section 6.1.

Note that the variables recording the number of cases with valid values on **A3** and the total number of employees participating in each workplace are derived for the purposes of assessing the extent to which the information provided by those employees that participated in the Survey can be taken to represent the wider workforce of which they are a part (see Section 6.3.3 for further details on this point).

Using syntax:

Having opened the Survey of Employees data file and ensured that the data is weighted (see Section 4.5.1), the syntax needed to produce the new, aggregated data file **Seq9ag.sav** is as follows.

```
aggregate outfile="d:\wers98\seq98ag.sav"
  /break=serno
  /avghrs=mean(a3)
  /avghrsok=nu(a3)
  /seqnum=nu .
```

Note that the original Survey of Employees data file remains as working data file, unless once replaces the new file name given on the `aggregate` command with an asterisk. As a result, the new data file **Seq98ag.sav** is not immediately available for analysis after completion of the command. Instead, the Survey of Employees data file must be closed and the new data file opened in its place.

We have given the unique workplace identifier **SERNO** as the 'break' variable on the `aggregate` command, so the new data file contains one record for each workplace with participating employees in the Survey of Employees data file. All of the subsequent variables are calculated across matching values of this variable.

The first new variable on the data file, **AVGHRS**, contains the mean number of hours worked by employees participating in the Survey of Employees, calculated within

---

workplace-level file can then manipulated with the use of the `vector` command. This alternative is not covered in this note, since the `aggregate` command should cover most users' needs.

each workplace over all cases where **A3** contains a valid response. The second variable, **AVGHRSOK**, contains the unweighted number of cases in which **A3** contains a valid response. The third variable, **SEQNUM**, holds the unweighted number of cases from each workplace that are present in the Survey of Employees dataset. This variable will necessarily have a minimum of 1 and, because of the sample design, a maximum of 25.

For a list of other useful functions that may be specified on the `aggregate` command, besides `mean`, users are referred to the SPSS on-line Users Manual.

Using the menu system:

1. Open the Survey of Employees data file **Seq98.sav** and ensure that the data is weighted (see Section 4.5.1).
2. From the 'Data' menu, choose the option labelled 'Aggregate…'.
3. From the list of variables on the left hand side of the new window, select **SERNO** and use the arrow button to transfer it into the list headed 'Break Variable(s):'.
4. The first new variable we wish to create will contain the mean number of hours worked by employees participating in the Survey of Employees, calculated within each workplace over all cases where **A3** contains a valid response. To do this, select the variable **A3** and use the lower of the two arrow buttons to transfer it into the list headed 'Aggregate Variable(s):'.
5. Clicking on the button labelled 'Name & Label' will allow you to alter the name and label of the new, aggregated variable which will, by default, be named **A3_1**. Change it to **AVGHRS** to better reflect the function of the new variable.
6. Clicking on the button labelled 'Function' would allow you to alter the function used in creating the new, aggregated variable. However, the default is 'mean', which is what we require and so it can be left as is.
7. To create a second new variable that counts the number of cases in which **A3** is missing (i.e. does not contain a valid response), again select **A3** in the list on the left hand side of the window and use the lower of the two arrow buttons to transfer it into the list headed 'Aggregate Variable(s):'.
8. Click on the button labelled 'Name & Label' and change the name of the variable from **A3_2** to **AVGHRSOK**.
9. Click on the button labelled 'Function' and change the function from 'Mean of values' to 'Number of cases'. We want the unweighted count of the number of cases with valid values on **A3**, so having checked the circle labelled 'Number of cases', we also check the box labelled 'Unweighted', leaving the box labelled 'Missing' unchecked. Click on the button labelled 'Continue' to return to the first window.
10. To set up the third new variable, which holds the number of cases from each workplace that are present in the Survey of Employees dataset, select **SERIAL** in the list on the left hand side of the window and use the lower of the two arrow buttons to transfer it into the list headed 'Aggregate Variable(s):'.
11. For consistency with the syntax given above, rename this variable **SEQNUM**.
12. Click on the button labelled 'Function' and change the function from 'Mean of values' to 'Number of cases'. We want the unweighted count of the number of employees from each workplace present in the data file, so check the box labelled 'Unweighted'. Click on the button labelled 'Continue' to return to the first window.

13. Finally, ensure that the option to 'Create new data file' is selected and change the name of this file as appropriate. In the syntax example, we named the file **d:\wers98\Seq98ag.sav**. Note that, in this case, the original Survey of Employees data file remains as working data file. As a result, the new data file – **Seq98ag.sav** - is not immediately available for analysis after completion of the command. Instead, the Survey of Employees data file must be closed and the new data file opened in its place. To make the new data file the working data file as the command is run, check the option labelled 'Replace working data file'.
14. Finally, clicking on the button labelled 'OK' will run the `aggregate` command and create the new, aggregated data file.

Creating additional new variables merely involved repeating Steps 4 to 6, changing the source variable, name and function as required.

6.3.2 Aggregating data from the Survey of Employees in STATA

The employee data can be aggregated at workplace level in STATA by using the `collapse` command to generate a workplace-level data file that contains summary information about the employees from that workplace who participated in the Survey of Employees (e.g. mean number of hours worked).

Suppose that we wished to create a workplace-level data file containing three summary data items from the Survey of Employees: first, the mean number of hours worked by the participating employees in each workplace; second, the number of employees giving a valid (non-missing) response to the question on hours; and third, the total number of participating employees in each workplace. This workplace-level data file could then be matched onto the Management or Worker Representative data files using the method outlined in Section 6.1.

Note that the variables recording the number of cases with valid values on **A3** and the total number of employees participating in each workplace are derived for the purposes of assessing the extent to which the information provided by those employees that participated in the Survey can be taken to represent the wider workforce of which they are a part (see Section 6.3.3 for further details on this point).

The `collapse` command takes the Survey of Employees data file and creates a new data file in which there is one record for each workplace. As with SPSS's `aggregate` command, `collapse` can create a range of summary data items containing, for example, the mean value of a particular variable for employees in that workplace, the minimum or maximum value amongst those employees, or the sum of all values amongst those employees.

However, unlike `aggregate`, `collapse` cannot directly compute unweighted numbers of cases from weighted data. We therefore need to incorporate an additional step, in which we create two dummy variables. The first will be used to count the number of employees in each workplace that gave a valid response at **A3**, and so takes the value of 1 in such cases and the value of 0 otherwise. The second dummy will be used to count the number of employees in each workplace that participated in the Survey of Employees, and so takes the value of 1 in all cases.

Using syntax:

Having opened the Survey of Employees data file and ensured that the data is weighted (see Section 4.5.2), the syntax needed to produce the aggregated data is as follows:

```
gen avghrchk=(a3<.)
gen avghrchk=(a3<.)
collapse (mean) avghrs=a3 (rawsum) avghrsok=avghrchk
     seqnum=seq [pw=empwt_nr], by (serno)
```

We have given the unique workplace identifier **SERNO** as the 'break' variable, so this aggregated data set contains one record for each workplace with participating employees in the Survey of Employees data file. All of the subsequent variables are calculated across matching values of this variable.

The first new variable on the data set, **AVGHRS**, contains the mean number of hours worked by employees participating in the Survey of Employees, calculated within each workplace over all cases where **A3** contains a valid response. The second variable, **AVGHRSOK**, contains the number of cases in which **A3** contains a valid response. The third variable, **SEQNUM**, holds the number of cases from each workplace that are present in the Survey of Employees data file. This variable will necessarily have a minimum of 1 and, because of the sample design, a maximum of 25.

The collapse command creates a new workplace-level data set that can be analysed immediately. However, the data set is only held in memory and is not saved by the procedure – a departure from the practice of the SPSS `aggregate` command. In STATA, the aggregated data set needs to be saved using the normal methods.

For a list of other useful functions that may be specified on the `collapse` command, besides `mean`, users are referred to the relevant entry in the STATA Reference Manual.

6.3.3 A note about the generalizability of aggregated data from the Survey of Employees

By deriving the variables **AVGHRSOK** and **SEQNUM** in Sections 6.3.1 and 6.3.1, we have hopefully hinted at the question of the generalizability of information that is obtained by aggregating data from the Survey of Employees. Two issues need to be addressed in the analysis of the data: response bias and precision.

Response bias:

If the aggregated data is biased in some way, it will not accurately characterize the population that it is expected to represent. Bias may be introduced into the aggregated data from two sources.

The first potential source of bias arises from employee non-response within the Survey of Employees. So, in any given workplace, if the response rate among

employees selected to participate in the Survey of Employees was less than 100 per cent, it is possible that those who responded may constitute a biased sample of the those that were selected.

One cannot formally assess whether there is any bias as one does not know the profile of those employees that were asked to participate in the Survey within each workplace. However, one can minimize the risk of such bias being present in aggregated data by only compiling aggregate measures in workplaces with relatively high response rates on the Survey of Employees. A response rate of 60 per cent would seem to be a reasonable benchmark. Applying this threshold means that, in workplaces with 25 or more employees (where 25 questionnaires were distributed), any aggregate workplace-level measure would need to be based on at least 15 employee records. In a workplace with only 10 employees (where all employees received a questionnaire), at least 6 must have returned their questionnaire. This '60 per cent' rule is the benchmark advocated by the team responsible for the employee survey within the Australian Workplace Industrial Relations Survey of 1995 (Morehead and Alexander, 1999: 550).

Applying such a rule necessarily means that one will be calculating aggregate measures for only a selection of workplaces that participated in the Survey of Employees. A survey response rate of at least 60 per cent was achieved in some 1,219 workplaces in WERS98. These workplaces represent 68 per cent of the 1,782 establishments that participated in the Survey of Employees and 56 per cent of the 2,191 that took part in the Cross-Section survey as a whole. Of course, some of the individual questions in the Survey of Employees have additional degrees of non-response, and so the number of workplaces passing the threshold will be lower for individual variables (hence the reason for deriving the two variables **SEQNUM** and **AVGHRSOK** in Sections 6.3.1 and 6.3.2).

We therefore need to consider whether any bias is introduced into the workplace-level sample that we will use in our analysis as a result of our exclusion of workplaces with SEQ response rates of less than 60 per cent. This is the second potential source of bias. In doing so, we should also consider whether any bias is introduced into our final, workplace-level sample as a result of workplace non-participation in the Survey of Employees. Even if we set no threshold on the number of employee responses needed to compile aggregate measures, and use all of the workplaces for which at least one employee returned a questionnaire, this sample of workplaces may still be unrepresentative of all workplaces covered by WERS98. The WERS98 Technical Report indicates that larger workplaces and those in certain industries, such as Hotels and Restaurants, were less likely to agree to participate in the Survey of Employees (Airey *et al.*, 1999: 61). Equally, there may be other workplace characteristics that were associated with management's refusal to participate.

One can assess the extent of any workplace-level bias by comparing the profile of those workplaces for which one has compiled aggregate measures with the profile of all workplaces participating in WERS98. If the profiles are appreciably different across a particular variable (e.g. workplace size), and that variable is associated in some way with value of the dependent variable you are estimating, then estimates based on the aggregated sample may not be fully representative of the whole. In such cases, adjustments may need to be made to your estimates to remove the bias. In

regression analysis, this is done through a two-stage estimation process using the Heckman procedure, whereby one first estimates the probability of a case featuring in the final sample and then incorporates the resulting selection term into a model of the dependent variable under investigation.

Precision:

If the responses are unbiased, one must still be concerned with the question of how precisely the employee data will represent the characteristics of the workforce as a whole within any particular establishment given that, in many cases, we have obtained data from only a fraction of the workforce. Below, we show the implications that different achieved sample sizes have on the precision of aggregated data from the Survey of Employees. We look first at dichotomous variables; then means or proportions.

Dichotomous variables:

Suppose that, in a workplace with 2,000 employees, 60 per cent are satisfied with their work. We wish to construct a dichotomous variable indicating whether at least half of the workforce are satisfied with their jobs. However, we have only surveyed 25 of the 2,000 employees. Furthermore, only 20 have returned the questionnaire and filled in the relevant questions on job satisfaction. Assuming that the 20 are an unbiased sample, what is the probability that our dichotomous variable, based on information from only 20 employees, will incorrectly indicate the balance of satisfaction in the workforce as a whole?

In this case, the answer is about 0.13. In other words, we can expect that we will incorrectly gauge the views of the majority in about 13% of all cases. This 'probability of error' can be calculated using the hypergeometric distribution (Hymans, 1967: 146-7). The hypergeometric distribution is similar to the binomial distribution but, whereas the binomial applies to cases that have been sampled *with* replacement, the hypergeometric applies to cases that have been sampled *without* replacement.

Unfortunately, this 'probability of error' is not a particularly easy statistic to calculate. Therefore, for illustrative purposes, we have provided a table that contains some calculations of this 'probability of error' for different sizes of workplace and valid sample. The table assumes that 60 per cent of the workforce possess the characteristic in question. In reality, this figure cannot be known. Suffice it to say that the 'probability of error' calculated by the hypergeometric distribution decreases rapidly as this 'population percentage' moves further away from 50 per cent, and vice versa.

Table 1 Percentage of workplaces in which a dichotomous variable based on SEQ returns can be expected to incorrectly indicate the characteristics of the majority of the workforce

If 60 per cent of the whole workforce possess the characteristic:

	Number of valid returns in SEQ dataset				
Size of workforce	**5**	**10**	**15**	**20**	**25**
10	26	0	-	-	-
25	30	11	10	0	0
50	31	14	17	7	7
100	31	15	19	10	12
500	32	16	21	12	15
1000	32	17	21	13	15
2000	32	17	21	13	15

The table shows that the probabilities of error in our particular variable for samples of 20 and 25 are broadly equivalent. However, in larger workplaces (100 or more employees), the likelihood of error does not differ greatly among samples of 10 or more employees. This is partly because our variable defines the workforce as 'majority satisfied' if 50 per cent or more of the sample are satisfied. As a result, in even-numbered samples (i.e. 10 or 20), the marginal cases (i.e. where 5 or 10 of the sample are satisfied, respectively) are accepted. If we wished to identify only those workplaces in which a strict majority were satisfied, these marginal cases would constitute errors and so the 'probability of error' would be greater. In the case of a sample of 10, the probabilities would be broadly equivalent to those in the N=5 column (i.e. around 30 per cent). In the case of a sample of 20, the probabilities would be broadly equivalent to those in the N=15 column (i.e. around 20 per cent).

In view of this latter point, there would appear to be an appreciable loss in precision through basing estimates on samples of 10 employees or less. It would seem that a sample of 15 might reasonably be set as a lower bound for compiling dichotomous variables, as it was in our discussion of bias above.

Means or proportions:

Users wishing to use the Survey of Employees data to calculate workplace-level means or proportions (e.g. proportion satisfied with their pay) should first bear in mind the large degree of uncertainty that will surround point estimates, particularly in larger workplaces where only a small proportion of the workforce have been surveyed.[16]

To give an illustrative example, analysis of the whole Survey of Employees data file shows that 36 per cent of all employees were satisfied with their pay (question **A10B**).

[16] This uncertainty disappears completely when all of the employees at the workplace have been surveyed and all have returned their questionnaires, as is the case in 21 of the 1,782 workplaces that participated in the Survey of Employees.

This estimate has a standard error of around 0.5 and, hence, a 95% confidence interval of around 2 per cent. However, within those 34 workplaces in which 25 employee questionnaires were returned, the standard error was more like 6, on average. This generates an average 95% confidence interval of around 25 per cent for the workplace-level estimate.

One must also remember that the confidence intervals will be wider in workplaces where a smaller proportion of the sampled employees have returned their questionnaires. The following table illustrates how a standard error increases as the sample size falls progressively below 25, all other things remaining constant.

Table 2 Relative increase in standard errors for estimates based on samples of less than 25 employees

Sample size	**20**	**15**	**10**	**5**
Increase in SE when compared with sample of 25	12%	29%	58%	124%

Low sample sizes are therefore a particular problem in respect of the reliability of workplace-level means and proportions based on data from the Survey of Employees. Returning to the example of satisfaction with pay, we find that the standard error of the estimate among workplaces with 20 returns was around 8, on average, and where 15 questionnaires were returned it was around 10. This increase broadly follows that suggested in the table.

The conclusion, therefore, is that one must be particularly careful when constructing workplace-level means or proportions from the Survey of Employees data, in cases where only a fraction of the workforce were asked to participate, even if all of the selected employees have returned their questionnaires.

6.4 Combining interview data with verbatim text

6.4.1 The spreadsheets of verbatim text

WERS98 is the first survey in the WIRS series for which verbatim answers given by respondents in the survey interviews have been made publicly available. This development, made possible by the use of Computer Assisted Personal Interviewing (CAPI), offers researchers a number of new opportunities.

1. Researchers may wish to search for particular types of answer, not separately identified by the Research Team's code frames. For example, one might wish to identify respondents with the job title 'Industrial Relations Manager'. This job title is combined with other titles on code 3 of the categorical variable **BTITLE2**, but relevant cases can be separately identified from the verbatim answers to the original open-ended question, **BTITLE**.
2. Alternatively, one may wish to compile a new code frame to be used in place of that developed by the WERS98 Research Team. This could be particularly useful if comparing results from WIRS90 and WERS98 in cases where the code frame for a particular question has been changed. **BTITLE2** again provides a good example.

3. Finally, researchers may wish to use textual analysis software, such as NU*DIST, to look for patterns in verbatim answers. This might prove fruitful with respect to the verbatims collected at question **D12** in the Survey of Employees, for example.

The verbatim answers are held in four **restricted access** Excel spreadsheets, as follows:

Cross-Section Management interview	MQOPEN.XLS
Cross-Section Worker Representative interview	WRQOPEN.XLS
Cross-Section Survey of Employees (**D12** only)	SEQOPEN.XLS
1998 Panel Survey interview	PQOPEN.XLS

The three files that derive from face-to-face interviews each contain verbatim responses to partially open questions, such as **AHEADOFF**, and fully open questions, such as **BTITLE**.

Note, however, that the answers contained in all four of the files have been anonymized in order to protect the confidentiality of respondents. This means that all references to organization names or individuals have been replaced by a string of xxxxx's.

6.4.2 How to export data from a spreadsheet for use in SPSS or STATA

Users following routes 1 or 2 from the previous section will need to match their numeric codes back onto the interview data before the new coding system can be used for analysis. The procedures required to do this are quite straightforward.

Using SPSS syntax:

Once you have recoded the verbatims in Excel, the spreadsheet page containing your new coding must first be saved as a single Excel 4.0 worksheet, since SPSS cannot read in spreadsheets created using Excel 5.0 or later. Having created this Excel 4.0 sheet, one can then use the `get translate` command to read the data into the SPSS Data Editor. The `get translate` command takes the following basic form:

```
get translate file='d:\wers98\sheet1.xls'
     /type=xls .
```

Here, `d:\wers98\sheet1.xls` is the Excel 4.0 worksheet; `type=xls` specifies that it is an Excel file.

The optional `fieldnames` subcommand can also be specified in cases where the first row of the spreadsheet contains column headings that we wish to use as variable names. Specifying `fieldnames` means that SPSS automatically names the new variables according to these column headings.

The `range` subcommand can be specified if we wish to import only a rectangular selection of data from the spreadsheet. So if the spreadsheet had the unique workplace **SERNO** in its first column, the new numeric code in the second column and original

codes and verbatim text in subsequent columns, we could use `range` to read in only the first two columns of information from the sheet.

If we were to specify both of these options, the `get translate` command would take the following form:

```
get translate file='d:\wers98\sheet1.xls'
     /type=xls
     /fieldnames
     /range=a1:b300.
```

Here, we import only the rectangular range of data from cell A1 to cell B300.

Having imported the data from the spreadsheet into the SPSS Data Editor, the data can be saved as an SPSS data file in the normal way. It can then be matched onto the main interview data using the match files command, as explained in Section 6.1.1 and 6.2.1.

Using the SPSS menu system:

1. Once you have recoded the verbatims in Excel, the spreadsheet page containing your new coding must first be saved as a single Excel 4.0 worksheet, since SPSS cannot read in spreadsheets created using Excel 5.0 or later.
2. In SPSS, select the option labelled 'Open' from the 'File' menu.
3. In the box labelled 'Files of type:' at the bottom of the 'Open File' window, select 'Excel (*.xls)' to display all Excel files. Select your new Excel 4.0 spreadsheet and click on the button labelled 'Open'.
4. A new window will appear labelled 'Opening File Options'.
   a) The 'Read variable names' box should be checked if the first row of the spreadsheet contains column headings that you wish to use as variable names. Checking the box means that SPSS will automatically name the new variables according to the text in each column heading.
   b) One can insert a range if one wishes to import only a rectangular selection of data from the spreadsheet. So if the spreadsheet had the unique workplace **SERNO** in its first column, the new numeric code in the second column and original codes and verbatim text in subsequent columns, we could use `range` to read in only the first two columns of information from the sheet. If the spreadsheet contained 300 rows, we would specify the range as A1:B300.
5. Click on the button labelled 'OK' to import the spreadsheet data into the Data Editor.

Having imported the data from the spreadsheet into the SPSS Data Editor, the data can be saved as an SPSS data file in the normal way. It can then be matched onto the main interview data using the match files command, as explained in Section 6.1.1 and 6.2.1.

Using STATA syntax:

Once you have recoded the verbatims in Excel, the spreadsheet page containing your new coding must first be saved as a tab or comma-delimited text file, since STATA

cannot read in Excel files directly. Having created this file (which is easily done using Excel's 'Save as' option), one can then use the `insheet` command to read the data into STATA. The `insheet` command takes the following basic form:

```
insheet using 'd:\wers98\sheet1.txt', names tab
```

if the file is tab-delimited, or

```
insheet using 'd:\wers98\sheet1.csv', names comma
```

if the file is comma-delimited.

The `names` sub-command tells STATA that the first row of the spreadsheet contains column headings that you wish to use as variable names. Inserting this sub-command means that STATA will automatically name the new variables according to the text in each column heading. If you do not wish STATA to do this, simply omit the `names` sub-command.

If you do specify the `names` sub-command, you may also read in just a selection of variables from the spreadsheet. To do this, simply list the variables between the words insheet and using, as in the following example:

```
insheet serno newvar using 'd:\wers98\sheet1.txt', names
     tab
```

Having imported the data from the spreadsheet into STATA, the data can be saved as a STATA data file in the normal way. It can then be matched onto the main interview data using the match files command, as explained in Section 6.2.1 and 6.2.2.

6.4.3 How to export data from SPSS or STATA and add it to a spreadsheet

Users following route 3 in the opening part of this section will probably wish to export additional data items from the survey data files and add them into the spreadsheets of verbatim answers. For example, when analysing the verbatims from **D12** in the Survey of Employees, it may be helpful to be able to refer to the employee's gender, age or other characteristics. To do this, users will need to write out a spreadsheet file from SPSS or STATA containing the required data items. Specific Excel functions can then be used to match these data items onto the relevant cases in the spreadsheet of verbatims. Each stage is outlined below.

Writing out a spreadsheet file from SPSS, using syntax:

First, one should create an SPSS data file containing the relevant data items. Note that the unique case identifier (**SERNO**, **SERIAL** or **SERNO2**, depending upon which data file is being used) should be the first item on the data file. The data file should also be sorted in ascending order of this variable. This SPSS data file can then be exported as an Excel 4.0 spreadsheet (`d:\wers98\sheet2.xls`) using the `save translate` command:

```
save translate outfile='d:\wers98\dataserv\check2.xls'
     /type=xls
     /fieldnames.
```

The optional `fieldnames` subcommand can also be specified when one wishes the variable names of the SPSS variables to be copied into the first row of the new spreadsheet as column headings.

Writing out a spreadsheet file from SPSS, using the menu system:

1. Create an SPSS data file containing the relevant data items. The unique case identifier (**SERNO**, **SERIAL** or **SERNO2**, depending upon which data file is being used) should be the first item on the data file. The data file should also be sorted in ascending order of this variable.
2. Choose the 'Save As…' option from the 'File' menu in SPSS.
3. In the box labelled 'Save as type:', choose 'Excel (*.xls)' and give the new file a name.
4. If one wishes the variable names of the SPSS variables to be copied into the first row of the new spreadsheet as column headings, check the box labelled 'Write variable names to spreadsheet'.
5. Click on the button labelled 'Save'.

An Excel 4.0 spreadsheet will be written out by SPSS.

Writing out a spreadsheet file from STATA, using syntax:

First, one should create a STATA data file containing the relevant data items. Note that the unique case identifier (**SERNO**, **SERIAL** or **SERNO2**, depending upon which data file is being used) should be the first item on the data file. The data file should also be sorted in ascending order of this variable. This STATA data file can then be exported as a tab-delimited, spreadsheet-style file (`d:\wers98\sheet2.txt`) using the `outsheet` command:

```
outsheet using "d:\wers98\sheet2.txt", nolabel
```

The `nolabel` option specifies that data values (rather than value labels) are written to the new file. One can also specify the `nonames` option if one doesn't want variable names to appear in first row of the new spreadsheet file.

This new file, `d:\wers98\sheet2.txt`, can be read into Excel as a tab-delimited file, and then saved as an Excel spreadsheet in the normal way.

Matching the data with the verbatims in Excel:

1. Open the spreadsheet containing the verbatim answers (the spreadsheet into which you wish to import the interview data). Sort the file in ascending order of the unique case identifier (**SERNO**, **SERIAL** or **SERNO2**, depending upon which data file is being used).
2. Create a blank column to hold the first item of data that you wish to import and insert a descriptive title in the first row.

3. In the second row of this new column, insert the following function, replacing each italicised argument with relevant values as described below. The function is:

   `=vlookup(`*`value_to_match, datafile_dimensions, data_col`*`)`

   where:

   value_to_match is the cell reference of the unique case identifier in the open spreadsheet (the one containing the verbatim)
   datafile_dimensions gives a full reference to the second data file and the range of cells within it that contain data
   data_col is the number of the column in this second spreadsheet that contains the data item that you wish to import.

   A completed function might look like this:

   `=VLOOKUP(A2,'d:\wers98\sheet2.xls'!$A$2:$C$300,2)`

   In this case, data from column 2 of the second spreadsheet will be imported into the cell containing the `vlookup` function, as long as a match can be found between the unique case identifier in the verbatims file (held in cell A2) and a value held in the first column of the second spreadsheet.

7. Acknowledging the use of the WERS98 data in publications

7.1 Acknowledgement and disclaimer

Users are reminded that the undertaking which is given to the Data Archive prior to receiving data from WERS98 requires them to acknowledge the roles of the both the original depositors and the Archive in any publication, whether printed, electronic or broadcast, based wholly or in part on WERS98 data. The suggested wording is as follows:

> "The author acknowledges the Department of Trade and Industry, the Economic and Social Research Council, the Advisory, Conciliation and Arbitration Service and the Policy Studies Institute as the originators of the 1998 Workplace Employee Relations Survey data, and the Data Archive at the University of Essex as the distributor of the data. None of these organizations bears any responsibility for the author's analysis and interpretations of the data."

Those using the 1990-98 Panel Survey data should replace the words '1998 Workplace Employee Relations Survey (WERS98) data' with '1990 Workplace Industrial Relations Survey data and the 1998 Workplace Employee Relations Survey (WERS98) data'.

7.2 Bibliographic citation

All works that use the data should also acknowledge their source by means of bibliographic citation. To ensure that such source attributions are captured for bibliographic indexes, citations should appear in either: a footnote; an endnote; or, if using the Harvard style of referencing, the reference list of publications. Those using the Harvard system of referencing should insert (Department of Trade and Industry, 1999) in the main body of the work at the point of first reference to the data. The appropriate wording to be used for the full citation is as follows:

> Department of Trade and Industry (1999) *Workplace Employee Relations Survey: Cross-Section, 1998* [computer file]. 4th ed. Colchester: The Data Archive [distributor], 22 December 1999. SN: 3955.

Or, if using the 1990-98 Panel Survey data file:

> Department of Trade and Industry (1999) *Workplace Employee Relations Survey, 1998: Panel Survey 1990-1998* [computer file], Colchester: The Data Archive [distributor], 20 December 1999, SN: 4026.

7.3 Depositing copies of publications and derived data sets

The same undertaking also requires the user to deposit with the Data Archive two copies of any published work or report based on WERS98 and one copy of any new data sets which have been derived from the source data.

8. The WIRS bibliography

The WERS98 Data Dissemination Service web site (found at: www.niesr.ac.uk/niesr/wers98) contains a bibliography of all known publications arising from the analysis of data from the WIRS series. This bibliography lists all of the publicly available papers, of which we are aware, that have made original use of the data from the Workplace Industrial Relations Surveys (WIRS) Series. This series includes the 1998 Workplace Employee Relations Survey, as well as previous Workplace Industrial Relations Surveys of 1980, 1984 and 1990. The bibliography includes references to the books containing the primary analyses from each survey as well as numerous sources of secondary analysis, including books, journal articles and working papers. Over 200 items are currently listed.

The bibliography will be regularly updated as new research is published using WERS98. However, we rely upon users to assist us in keeping the bibliography up to date. We therefore request all users to please notify the Data Dissemination Service (by post or e-mail) of any new publications that use data from the WIRS series, as well as the publication of new versions of papers already listed in the bibliography (e.g. the progression of a working paper into a journal).

Appendix A: List of WERS98 Data Files and Documentation

Tables 1 and 2 in this Appendix list each of the WERS98 data files that are currently available. Table 3 lists additional data files that are to be made available in due course by the WERS98 Data Dissemination Service. Tables 4 to 6 list the various pieces of documentation that are currently available, or will be made available in future.

Note: In Tables 1 and 2 an asterisk (*) in place of a filename suffix (e.g. **Mq98fin.***) indicates that the suffix is dependent upon the format of the file. In the case of some data formats, notably SAS, the program files used to generate the data file are provided to the user by the Data Archive along with the data files themselves. The WERS98 data files are currently available in the following formats:

SPSS (portable files)	.POR	Data file
STATA	.DTA	Data file
SAS for Windows	.SD2	Data file
	.SAS	Program file
SAS for Unix	.SSD01	Data file
	.SAS	Program file
ASCII Comma-delimited	.CSV	Data file
ASCII Tab-delimited	.DAT	Data file
	.LST	Dictionary file

Files with the suffix .XLS are in Microsoft Excel format.

Table 1: Cross–Section Data Files currently available from the Data Archive:

Data File	Description	Number of Cases
General Release		
MQ98FIN.*	Contains data from the interview with the management respondent in the WERS98 cross-section. Also includes data from the Employee Profile Questionnaire (EPQ).	2,191
WRQ98.*	Contains data from the interview with a nominated worker representative in the WERS98 cross-section.	918
SEQ98.*	Contains data from a random sample of 25 employees in workplaces that participated in the WERS98 cross-section.	28,215[a]
Restricted Release		
REGION.*	Data file of the regional identifiers: Government Office Region (GOR) and Standard Statistical Region (SSR) of the workplace.	2,191
LOCAL98.*	Contains information on unemployment rates and vacancies (average number unfilled and rates) by Government Office Region, Standard Statistical Region and Travel To Work Area (TTWA). All TTWA rates are banded.	2,191
MQ98_SIC.*	Contains SIC 92 at group level (4/5 digit).	2,191
SAMPLE98.*	Contains the variables that were used in the sampling for the 1998 Cross-Section survey (stratifiers and sampling fractions). Also contains a variable indicating the type of data available for each productive workplace in the Cross-Section Survey.	2,191
MQOPEN.XLS	Contains verbatim responses from open-ended questions in the interview with the management respondent in the WERS98 cross-section.	2,191
WRQOPEN.XLS	Contains verbatim responses from open-ended questions in the interview with the nominated worker representative in the WERS98 cross-section.	865
SEQOPEN.XLS	Contains verbatim responses from the open-ended question **D12** in the self-completion questionnaires distributed at workplaces participating in the WERS98 cross-section.	28,215[a]

[a] The data file that is available from the Data Archive actually contains 28,240 cases, but 25 of these arise from an establishment that did not yield a productive workplace interview (SERNO: 13068). See the volume of *Variable Notes* relating to **Seq98.*** for further details.

Table 2: Panel Data Files currently available from the Data Archive:

Data File	Description	Number of Cases
General Release		
PQ_9098.*	Contains data from the interviews with management respondents to the WERS98 Panel Survey. Also contains complete data from the interviews conducted at the same workplace in 1990 (as part of the 1990 Workplace Industrial Relations Survey).	846
PQ_98OUT.*	Contains data on the 1998 survival status of all 2061 workplaces interviewed as part of the 1990 Workplace Industrial Relations Survey, together with complete data from the 1990 interviews.	2,061
Restricted Release		
PQ_90REG.*	Standard Statistical Region and local unemployment rates at the time of the 1990 interview, for all cases contained in PQ_98OUT.POR or PQ_9098.POR.	2,061
PQ_98REG.*	Standard Statistical Region, Government Office Region and local unemployment and vacancy rates at the time of the 1998 interview, for all cases contained in PQ_9098.POR.	846
PQ_90SIC.*	Standard Industrial Classification (1980 Classification, Activity level) at the time of the 1990 interview, for all cases contained in PQ_98OUT.POR or PQ_9098.POR.	2,061
PQ_98SIC.*	Standard Industrial Classification (1980 Classification, Activity level) at the time of the 1998 interview, for all cases contained in PQ_9098.POR.	846
PQOPEN.XLS	Contains verbatim responses from open-ended questions in the 1998 interview of the WERS98 Panel Survey.	846

Table 3: Further Data Files to be made available by the WERS98 Data Dissemination Service:

Data File	Description	Number of Cases
General Release		
TIMESER.*	Contains consistently defined variables (where possible) for all data items that are present in the 1998 Cross-Section and at least one previous Cross-Section survey in the WIRS series.	8,049
MQ98DVS.*	Derived variables based on **Mq98fin.***	2,191
WRQ98DVS.*	Derived variables based on **Wrq98.***	918
SEQ98DVS.*	Derived variables based on **Seq98.***	28,240
PQ9098DV.*	Derived variables based on **PQ_9098.***	846
LEAVE90.*	Dataset of workplaces leaving the survey population between 1990 and 1998 (as used in 'All Change at Work?')	382
JOIN98.*	Dataset of workplaces joining the survey population between 1990 and 1998 (as used in 'All Change at Work?')	390
EXTDATA.*	Additional data items to be derived from external sources	2,191

Table 4: Components of the WERS98 User Guide:

Note: Available from the Data Archive and also on the Data Dissemination Service web-site

Document Description	Original Filename	Data Archive PDF File Containing Document	Data Dissemination Service PDF file
Introduction to WERS98.	INTRO.DOC	A3955UAB.PDF	Intro.pdf
Volume 1: Survey in Transition: A Guide to the design of WERS98.	VOLUME1.DOC		Survtran.pdf
Volume 2: WERS98 Technical Report.	VOLUME2.DOC		Tech_rep.pdf
Volume 7: WERS98 Interviewer Training Manual.	VOLUME 7.DOC		Training.pdf
Volume 3 Part A: Management Questionnaire.	VOLUME3A.DOC	A3955QAB.PDF	Mqver1_2.pdf
Employee Profile Questionnaire.	EPQ.PDF		Epqname.pdf
Volume 3 Part B: Worker Representative Questionnaire.	VOLUME3B.DOC		Wrqv1_3.pdf
Volume 3 Part C: Survey of Employees Questionnaire.	EMPLOYEE.PDF		Employee.pdf
Volume 5 Part A: Code Book for Cross-Section Datasets.	VOLUME5A.DOC	A3955CAB.PDF	Cbookv32.pdf
Additional Codes for the Cross-Section.	Not part of original User Guide		Addcodes.pdf
Volume 5 Part B: Instructions for Editing the Cross-Section Datasets.	VOLUME5B.DOC		Mqedit.pdf
Volume 5 Part C: Editing Instructions for The Employee Survey.	VOLUME5C.DOC		Seqedit.pdf
Basic Workforce Data Sheet.	BWDSNAME.PDF	A4026QAB.PDF	Bwdsname.pdf
Volume 4: The Panel Questionnaire.	PQ_Q12.DOC		Pq_q12.pdf
Volume 6 Part A: Code Book for Panel Dataset.	PQ_COD12.DOC	A4026CAB.PDF	Pq_cod12.pdf
Volume 6 Part B: Editing Instructions for the Panel Dataset.	PQ_ED.DOC		Pq_ed.pdf

Table 5: Additional documentation made available by the WERS98 Data Dissemination Service

Note: Available to download from the WERS98 Data Dissemination Service web-site. Each of the Notes is accompanied by a syntax file, also available from the web-site.

Document Description	Data Dissemination Service PDF file
Notes to Accompany the Management Dataset and Questionnaire	Mqnotes.pdf
Notes to Accompany the Worker Representative Dataset and Questionnaire	Wrqnotes.pdf
Notes to Accompany the Survey of Employees Dataset and Questionnaire	Seqnotes.pdf
Guide to Analysis of WERS98	Guide.pdf

Table 6: Further components of the WERS98 User Guide yet to be made available:

Document Description	Original Filename	Data Archive PDF File Containing Document	Data Dissemination Service PDF file
Volume 8: Documentation of Derived Variables from the Cross-Section Datasets.	To be confirmed		
Volume 9: Documentation of Derived Variables from the Panel Datasets.	To be confirmed		
Volume 10: A Guide to Using the WERS 80-98 Longitudinal Datasets.	To be confirmed		

Table 7: Further documentation to be made available by the Data Dissemination Service:

Document Description	Data Dissemination Service PDF file
Notes to Accompany the 1990-98 Panel Dataset and Questionnaire	Pqnotes.pdf

Appendix B: Contacting the Data Archive:

The contact details of the Data Archive are as follows:

Address: The Data Archive
University of Essex
Wivenhoe Park
Colchester
Essex
CO4 3SQ

Telephone: 01206 872001 (General Enquiries)

E-mail: archive@essex.ac.uk
Web-site: www.data-archive.ac.uk

Information on each of the WERS98 data files can be found in the on-line BIRON catalogue at the Data Archive. The Data Archive Study Numbers that are needed to find information on WERS98 through BIRON's search engine are:

3955 for the 1998 Cross-Section Survey.

4026 for the 1990-98 Panel Survey.

Study Number **33176** will produce details on all the surveys in the WIRS series.

The BIRON catalogue provides access to on-line versions of the documents that comprise the WERS98 User Guide. These documents are also available on the WERS98 Data Dissemination Service web-site.

Appendix C: Institutions providing short courses on the analysis of survey data using SPSS or STATA

Centre for Applied Social Surveys (CASS)

CASS is an ESRC Resource Centre hosted by National Centre for Social Research and the University of Southampton, with the University of Surrey. Courses are held at various locations around the UK.

Contact details:

Centre for Applied Social Surveys (CASS)
Department of Social Statistics
University of Southampton
Southampton SO17 1BJ
Tel: +44 (0)23 8059 3048
Fax: +44 (0)23 8059 3846
Email: cass@socsci.soton.ac.uk
URL: http://www.socstats.soton.ac.uk/cass/courses.html

The National Centre for Social Research and the University of Surrey also hold courses at their own institutions (see below).

National Centre for Social Research

The Survey Methods Centre at the National Centre for Social Research contributes to the running of courses at the Centre for Applied Social Surveys, but also runs its own internal courses for staff at the National Centre and the Office for National Statistics. These courses are now available to a wider audience.

Contact details:

Survey Methods Centre
National Centre for Social Research
35 Northampton Square
London EC1V 0AX
Tel: +44 (0)171 250 1866
URL: http://www.natcen.ac.uk

Department of Sociology, University of Surrey

The Department runs practical courses, taught by staff from the University's social research methods centre. Courses can also be run for a group, either at the University of Surrey or off-site.

Contact details:

Department of Sociology
University of Surrey
Guildford GU2 5XH
Tel: +44 (0)1483 259365
Fax: +44 (0)1483 259551
E-mail: short.courses@soc.surrey.ac.uk
URL: http://www.soc.surrey.ac.uk/daycourses/dcindex.html

SPSS UK Ltd.

SPSS UK Ltd. also offers short courses in the use of its software. The focus of these courses is more on the functionality of SPSS, rather than the principles of survey analysis. These courses can be considerably more expensive than those offered by academic institutions.

Contact details in the UK:

SPSS UK Ltd.
1st Floor
St. Andrew's House
West Street
Woking
Surrey GU21 1EB
Telephone: +44.1483.719200
Fax: +44.1483.719290
E-mail: training@spss.co.uk
URL: http://www.spss.com.uk/training.html

Outside the UK, see URL: http://www.spss.com/training/home.cfm

STATA Corporation

STATA offer course from introductory to advanced level that are administered via the Internet and E-mail. As with the courses offered by SPSS, the focus is on the functionality of the software. However, the courses are very reasonably priced.

For further information, consult the Netcourse page on the STATA web-site at the following address: http://www.stata.com/info/products/netcourse/

The official distributor of STATA in the UK is Timberlake Consultants. They also plan to begin running training courses via the Internet in 2000.

Contact details:

Timberlake Consultants Ltd
Unit B3 Broomsleigh Business Park
Worsley Bridge Road
London SE26 5BN
Telephone: +44 (0)208 697 3377

Fax: +44 (0)208 697 3388
E-mail: Info@timberlake.co.uk
URL: http://www.timberlake.co.uk

Appendix D: Contact details for the WERS98 Data Dissemination Service

The contact details of the WERS98 Data Dissemination Service are as follows:

Address: WERS98 Data Dissemination Service
c/o Simon Kirby
National Institute of Economic and Social Research
2 Dean Trench Street
Smith Square
London SW1P 3HE

E-mail: wers98@niesr.ac.uk
Web-site: http://www.niesr.ac.uk/niesr/wers98

Telephone: 020 7654 1902 (Direct line)

If you have any queries concerning WERS98 please do not hesitate to contact us. However, before doing so, please help us and other users by ensuring that the answer is not already provided in this *Guide to Analysis*, in the volumes of *Variable Notes* or on our web-site.

We would prefer, where possible, to receive queries by e-mail, which we aim to answer within three working days.

Appendix E: Output from the SPSS Tables module

Example 1

(EANYEMP + BASE1) BY (NEMPSIZE + TOTAL1)

		Size of establishment						
		0 10 thru 24 employees	1 25 to 49 employees	2 50 to 99 employees	3 100 to 199 employees	4 200 to 499 employees	5 500 or more employees	All w/places
Any employees who are a member of a trade union?	1 Yes	40	46	53	67	78	86	47
	2 No	60	54	47	33	22	14	53
Base	Weighted	1095	575	274	134	84	29	2191
	Unweighted	262	396	393	387	456	297	2191

Example 2

First table as in Example 1. Second table:

(APHRAS01 + BASE1) BY (ASTATUS + TOTAL1)

		How would you describe the formal status of this workplace / organisation?			
		1 Private sector company - PLC	2 Private sector - other	3 Public sector	All w/places
We frequently ask employees at our workplace to help us in ways not specified in their job?	1 Strongly agree	11	17	12	14
	2 Agree	43	33	39	38
	3 Neither agree nor disagree	11	10	15	11
	4 Disagree	31	34	29	32
	5 Strongly disagree	5	5	5	5
	9 Dont know			0	0
Base	Weighted	640	1007	544	2191
	Unweighted	834	680	677	2191

Example 3

(ZABSENCE + BASE1) BY (ASTATUS + TOTAL1)

		How would you describe the formal status of this workplace / organisation?			
		1 Private sector company - PLC	2 Private sector - other	3 Public sector	All w/places
Over the last twelve months what per cent?	Mean	4.4	4.2	5.4	4.6
Base	weighted	518	856	456	1831
	unweighted	693	548	544	1785

Example 4

(ZABSENCE + BASE1) BY (ASTATUS + TOTAL1)

		How would you describe the formal status of this workplace / organisation?			
		1 Private sector company - PLC	2 Private sector - other	3 Public sector	All w/places
Over the last twelve months what per cent?	Mean	4.5	4.1	5.0	4.5
Base	weighted	42503	37561	34052	114117
	unweighted	693	548	544	1785

Example 5

(BYOURJ + BASE1) BY (NEMPSIZE + TOTAL1)

		Size of establishment						
		0 10 thru 24 employees	1 25 to 49 employees	2 50 to 99 employees	3 100 to 199 employees	4 200 to 499 employees	5 500 or more employees	All w/places
Work responsibilities of respondent and their subordinates	1 Pay or conditions of employment	79	73	78	81	86	93	78
	2 Recruitment or selection of employees	93	93	89	93	94	89	93
	3 Training of employees	89	87	83	87	85	74	87
	4 Systems of payment	55	53	56	56	60	62	55
	5 Handling grievances	92	91	92	97	97	96	92
	6 Staffing or manpower planning	87	88	88	86	90	86	87
	7 Equal opportunities	87	91	85	90	95	95	88
	8 Health and safety	84	86	79	80	76	58	83
	9 Performance appraisals	82	80	86	83	85	85	82
	10 None of these		1	1	0			0
Base	Weighted	1095	575	274	134	84	29	2191
	Unweighted	262	396	393	387	456	297	2191

References

Airey C, Hales J, Hamilton R, McKernan A and Purdon S (1999) *The Workplace Employee Relations Survey (WERS) 1997-8: Technical Report (cross-section and panel surveys)*, London: National Centre for Social Research.

Brick J and Morganstein D (n.d.) "Analysis of complex samples using replication", SPSS White Paper, mimeo. Available on-line at: http://www.spss.com/cool/papers/white2b.htm [verified: 10/4/00].

Haltiwanger J, Lane J, Speltzer J, Theeuwes J and Troske K (eds.) (1999) *The Creation and Analysis of Employer-Employee Matched Data*, Amsterdam: Elsevier.

Hedges B (1978) "Sampling", in G Hoinville, R Jowell et al., *Survey Research Practice*, London: Heinemann.

Hymans S (1967) *Probability Theory with Applications to Econometrics and Decision-Making*, Englewood Cliffs, New Jersey: Prentice Hall.

Kish L (1965) *Survey Sampling*, New York: Wiley.

Millward N, Forth J and Bryson A (1999) "Changes in employment relations, 1980-1998", in M Cully, S Woodland, A O'Reilly and G Dix, *Britain at Work: As Depicted by the 1998 Workplace Employee Relations Survey*, London: Routledge.

Morehead A and Alexander M (1999) "The 1995 Australian Workplace Industrial Relations Survey", in J Haltiwanger, J Lane, J Speltzer, J Theeuwes and K Troske (eds.) (1999) *The Creation and Analysis of Employer-Employee Matched Data*, Amsterdam: Elsevier.

Pfefferman D (1996) "The use of sampling weights for survey data analysis", *Statistical Methods in Medical Research*,

Rao J and Thomas D (1989) "Chi-squared tests for contingency tables", in C Skinner, D Holt and T Smith (eds.) *Analysis of Complex Surveys*, Chichester: John Wiley and Sons.

Rust K (1985) "Variance estimation for complex estimators in sample surveys", *Journal of Official Statistics*, 1, 4: 381-97.

Skinner C (1989a) "Introduction to Part A", in C Skinner, D Holt and T Smith (eds.) *Analysis of Complex Surveys*, Chichester: John Wiley and Sons.

Skinner C (1989b) "Domain means, regression and multivariate analysis", in C Skinner, D Holt and T Smith (eds.) *Analysis of Complex Surveys*, Chichester: John Wiley and Sons.

Skinner C (1997) “The use of sampling weights in the regression analysis of WIRS data”, University of Southampton, mimeo.

Sribney W (1998) “Two-way contingency tables for survey or clustered data”, *Stata Technical Bulletin*, 45: 33-49.

Wolter K (1985) *Introduction to Variance Estimation*, New York: Springer-Verlag.